

'Big science' and public trust

Viewing the budget crisis enveloping Europe's leading particle-accelerator lab, politicians may conclude that high-energy physicists cannot be trusted to manage billions of dollars of public money. This perception must be reversed — and fast.

The first collision at the Large Hadron Collider (LHC) has occurred, somewhat earlier than expected. But it is not high-speed protons that have met head-on. Instead, projected cost overruns of several hundred million dollars have careered into the proud record held by CERN, the European particle physics laboratory near Geneva, for having previously delivered such projects on time and on budget.

Most worrying is the suggestion that poor financial management and inadequate accounting procedures are at least partly to blame (see page 557). In their defence, CERN's managers say that the true cost of developing and installing the required magnets and other technology could not have been fully predicted at the outset. In many respects, the laboratory was taking a step into the unknown, they argue. CERN also had no choice but to start working on the LHC without the sort of contingency funds that are often built into projects of this size — its member states were never prepared to stump up the necessary cash.

But CERN has so far made no attempt to absorb the cost of developing the prototype magnets for the LHC into the lab's central budget, as was intended. And while its director-general, Luciano Maiani, argues that it has only recently been possible to carry out a meaningful cost review, to have explicitly denied rumours of significant cash problems as recently as June is inexcusable.

Worse still, the lab then issued a self-congratulatory press release, entitled "1754 days to the LHC and counting!". No wonder representatives from CERN's member states feel aggrieved. CERN's decision

to set up an 'office for spending control' is a welcome move. But it is a belated one, and amounts to a tacit admission of faults in the financial management of the LHC project.

Many observers will question why such a prestigious horse was allowed to bolt before CERN installed, let alone closed, its stable door. This office needs to quickly demonstrate its ability to get the LHC project back on track, because more than CERN's reputation is riding on this particular horse.

The international high-energy physics community is now beginning to talk up its next big project — an enormous linear electron-positron collider. One reading of the current situation is that CERN's current troubles make it more likely that the next machine will be built in the United States. Certainly, after the débâcle of the Superconducting Supercollider, cancelled by the US Congress in 1993 amid spiralling costs, some US physicists — although not those who are heavily engaged at CERN — may take a grim pleasure in learning that their European counterparts are not immune to budget problems.

Given a gloomy economic climate, however, CERN's woes may convince the world's politicians that they have better things to do with public money than to give it to high-energy physicists. The scientific case for the next-generation linear collider may be strong, but politicians will need much convincing to commit money to it.

If the political fallout from the LHC crisis is not to lead to a global reluctance to fund such projects, CERN must first acknowledge that it has made mistakes, and then demonstrate that it has learnt from them. ■

Bangladesh's poisoned wells

Scientists and government officials bear collective responsibility for an unfolding tragedy.

The people of Bangladesh have endured flood, famine and disease. Now they must contend with a mass outbreak of arsenic poisoning that, according to one expert, makes the Chernobyl disaster "look like a Sunday-school picnic". The irony is that this catastrophe is the direct result of well-intentioned attempts to provide the nation with safe water for drinking and irrigation.

Up to 75 million Bangladeshis are at risk. A British court will soon be asked to decide whether one organization, the British Geological Survey, can be held accountable for its role in the crisis (see page 556). The case holds the prospect of hundreds — maybe thousands — of victims gaining compensation. But some experts fear that the threat of legal liability may in future deter Western hydrogeologists from working in the developing world.

That would compound the tragedy, because the expertise of Western scientists is still urgently needed. Wells will continue to be sunk across the Ganges delta, and without further research into the causes and extent of arsenic contamination, they may continue to draw from bodies of tainted groundwater.

Western hydrogeologists also have a moral obligation to the people of Bangladesh. Together with government officials and local

scientific experts, they must accept collective responsibility for the long delay in recognizing the problem.

When the first wells were sunk in Bangladesh in the 1970s, nobody suspected that arsenic could contaminate groundwater in river-plain sediments. Reports of arsenic poisoning related to wells sunk into such deposits began to emerge from West Bengal, across the Indian border, in 1983. Yet the first major conference on arsenic poisoning in the region and subsequent attention in the international scientific literature did not follow for more than a decade.

Government officials in Bangladesh and West Bengal — keen not to shoulder the burden of yet another public-health crisis in full view of the international community — had a hand in this. Warnings from local public-health experts in the mid-1980s were ignored. But both Western scientists and their colleagues on the Indian subcontinent must ask themselves whether they could not have done more to ensure that news of this public-health disaster emerged more quickly.

By accepting some of the blame, the scientific community may make it easier for local officials to take their share of the responsibility — and allow everyone to work towards preventing further loss of life. ■





Birthday honours
Alfred Nobel's prizes
celebrate their
centenary
p553–554



Balls of ire
Transgenic cotton
found growing
wild in India
p555



Poisoned chalice
Bangladeshis
prepare lawsuit over
arsenic in water
p556



Collision course
Cost overruns turn
spotlight on CERN's
management
p557

Trio united by division as cell cycle clinches centenary Nobel

Bernd Pulverer, London

The 2001 Nobels mark the centenary of the most prestigious prizes in science. So it is fitting that this year's award for physiology or medicine goes to three biologists who revealed the mysteries of a process central to the growth, development and maintenance of all living organisms: cell division.

Leland Hartwell, president and director of the Fred Hutchinson Cancer Research Center in Seattle, Paul Nurse, director-general of the Imperial Cancer Research Fund (ICRF) in London, and Tim Hunt, also of the ICRF, have been honoured for their contributions to our understanding of the cell cycle.

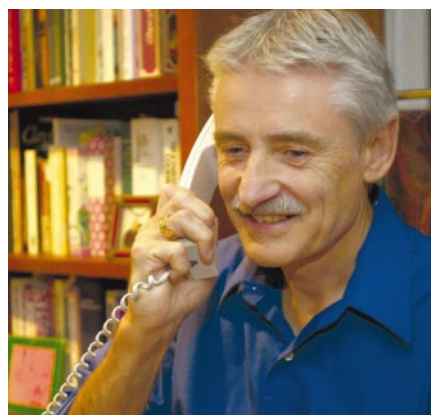
The molecular machinery underlying the cycle of cell growth and division remained obscure until the early 1970s, when Hartwell's pioneering work gave the first glimpses of the processes involved.

Working with baker's yeast, *Saccharomyces cerevisiae*, Hartwell set out to identify the genes involved in regulating the cell cycle. In a series of papers, he isolated yeast mutants that stopped at certain points in the cell cycle if the temperature was changed.

Hartwell identified more than 100 cell-division-cycle (CDC) genes, including one, dubbed *CDC28* or *start*, that seemed to initiate the copying of a cell's DNA, the step that precedes cell division. He went on to show that there was a striking interdependence between the events controlled by the various CDC genes—one would have to be completed before the subsequent stages could begin.

Many researchers who today work on the cell cycle say they were inspired to enter the field by Hartwell's pioneering studies. "Without his fundamental insight, there would be no such thing as the cell-cycle field," says Jim Roberts, also at the Hutchinson centre.

By studying the sensitivity of yeast cells to irradiation, Hartwell in the late 1980s introduced the concept of 'checkpoints', and identified the first yeast genes involved. Checkpoints are control points at which regulatory proteins halt the cell cycle upon detecting damage to DNA, or other abnormalities, allowing repair before the cycle resumes. In their absence, cells would be



Cell celebrators: centenary laureates Leland Hartwell (left), Paul Nurse (centre) and Tim Hunt.



prone to genetic damage. Indeed, many forms of cancer have since been linked to problems with checkpoint control.

Shortly after Hartwell began identifying CDC genes, Nurse started working on another species of yeast, *Schizosaccharomyces pombe*. He identified mutants involved in regulating the cell cycle, and later showed that one, labelled *cdc2*, was functionally equivalent to *start*. *CDC28* and *cdc2* have since been shown to encode enzymes belonging to the family of cyclin-dependent kinases (CDKs).

Phenomenal success

In 1987, Nurse isolated the corresponding gene in human cells, which encodes an enzyme later named CDK1. This, says Nurse, confirmed that biologists working on the cell cycle in yeast were identifying fundamental phenomena conserved by evolution across a wide range of organisms—and validated the study of model organisms to gain insights into human biology. "The same gene controlled cell division in yeast and humans and hence everything in between," he says.

Hunt's contribution was the discovery, from studies of fertilized eggs of frogs and sea urchins, of a new class of proteins that he called cyclins. The activity of CDKs was later found to depend on their association with these proteins.

In 1983, Hunt showed that the cyclins disappeared just before cell division, and

immediately realized that this was a key finding. "I knew I had made an important discovery," he says. Protein degradation has since emerged as an important general phenomenon in cell-cycle regulation.

Contacted by *Nature* on 8 October shortly after receiving the phone call from Stockholm, Nurse and Hunt were in ecstatic mood, with champagne corks popping in the background. At a press conference later in the day, Nurse admitted that the news had him "running around like a headless chicken". Hunt added that it only sunk in after he had checked the citation on the Nobel website.

But as is often the case, the rule that restricts Nobel awards to three winners (see pages 560–564) means that other pivotal figures go unrewarded. "If anyone will be slightly disappointed, it will be Yoshio Masui," Nurse told *Nature*. In the 1970s, Masui, now a professor emeritus at the University of Toronto, found that an extract from frog eggs, called maturation-promoting factor, readied the eggs for their first cell division. This was later found to be composed of the CDK encoded by *cdc2* and a cyclin.

But researchers in the field agree that Hartwell, Nurse and Hunt are worthy winners. "I call Hartwell the 'father', Nurse the 'great preacher' and Hunt the 'artist' of cell division," says Mitsuhiro Yanagida of Kyoto University in Japan.

Additional reporting by Erica Klarreich.

www.nobel.se

Cool atoms make physics prize matter

Philip Ball

Wolfgang Ketterle of the Massachusetts Institute of Technology (MIT) and Carl Wieman and Eric Cornell of JILA, an interdisciplinary research centre in Boulder, Colorado, have won this year's Nobel Prize in Physics for their work in making and understanding Bose-Einstein condensates (BECs).

This new form of matter, a strange state in which a group of atoms behaves as a single particle, was first created in 1995 by Wieman and Cornell by cooling atoms of rubidium to within less than a millionth of a degree of absolute zero. Ketterle's group at MIT managed to make a condensate only months later.

The theoretical existence of BECs was first proposed by Albert Einstein in 1924, building on work by the Indian physicist Satyendra Nath Bose. All particles are either bosons or fermions. Quantum theory says that no two fermions can occupy the same quantum state, but that any number of bosons can, in principle, exist in the same state. Einstein predicted that, at very low



Winners: Eric Cornell and Carl Wieman, who made the first condensate, and Wolfgang Ketterle (right).



temperatures, all the particles in a sample of bosons — such as the rubidium atoms used by the JILA team — should fall into the same state and act as a single particle, later termed a BEC.

The realization of BECs was long-awaited, and Wieman and Cornell's 1995 paper (see *Science* **269**, 198–201; 1995) caused a storm in the world of fundamental physics. Ketterle and his colleagues at MIT have subsequently clarified many of the properties of condensates. Together, the winners' work has made possible a host of other experiments that probe fundamental aspects of quantum theory.

Few prizes have been as widely predicted as this year's physics award. The original 1995 discovery was thought by many at the time to be worthy of a Nobel prize. Some physicists suspect that the physics prize in 1997, awarded to the pioneers of the atom-trapping technologies that Wieman, Cornell and Ketterle used to cool their atoms — while undoubtedly deserved in its own right — was also given to prepare the path for honouring the breakthroughs at JILA and MIT.

Research into BECs is still highly active. Jakob Reichel and colleagues at the Ludwig-Maximilians University in Munich, for instance, have recently created an 'atom chip' in which electromagnetic fields control the movement of a condensate hovering above an electronic circuit (see *Nature* **413**, 498–501; 2001). BECs are also allowing the investigations of imperfectly understood quantum phenomena such as entanglement (see *Nature* **413**, 400–403; 2001). Entanglement will be one of the key features of future quantum computers.

Applications of the work are currently distant, although several possibilities exist. "I'm sure fascinating applications will happen," says Claude Cohen-Tannoudji of the École Normale Supérieure in Paris, one of the 1997 laureates. He cites the example of atom lasers, in which a beam of atoms from a condensate is used to build high-precision structures. ■

The Nobel chemistry prize was announced after *Nature* went to press. For coverage, see

♦ www.nature.com

Ig Nobels raise a welcome smile

Steve Nadis, Boston

The organizers of the Ig Nobel prize ceremony, held to reward research "that cannot or should not be reproduced", faced a difficult decision this year. Should this light-hearted event be cancelled in view of the atrocities of 11 September? After some deliberation, the show went on — to the approval of a raucous crowd of 1,200 who gathered on 4 October in Harvard University's Sanders Theatre.

"There was a period of time shortly after the terrorist attack when it was difficult to think about things of consequence, let alone something like this," says Marc Abrahams, editor of the *Annals of Improbable Research* and the Ig Nobels' master of ceremonies.

"But I heard from a lot of people who wanted us to do it. I was both surprised and happy."

Nobel laureates who were asked to take part had mixed feelings. "I personally cannot muster any great enthusiasm for the Igs under the circumstances," Boston University physicist Sheldon Glashow told *Nature*. But other laureates, including Harvard chemist Dudley Herschbach, elected to participate. "I don't think it's disrespectful to move on and treat people to some fun," he says.

In the event, the tone was anything but sombre. The cast of Nobel laureates engaged in the usual shenanigans, singing in a première of an opera, *The Wedding Complex*, that culminated in the 60-second wedding of

two geologists, Lisa Danielson and Will Stefanov of Arizona State University.

As for the main act, Peter Barss of McGill University in Canada won the medicine prize for studying injuries caused by falling coconuts, while Chittaranjan Andrade and B. S. Srihari of the National Institute of Mental Health and Neuroscience in Bangalore snared the public-health prize for discovering that nose-picking is a common activity among adolescents. The psychology prize went to Lawrence Sherman of Miami University in Ohio for his ecological study of glee in preschool children, and Buck Weimer of Pueblo, Colorado, won the biology prize for inventing airtight underwear fitted with charcoal filters to remove foul-smelling gases.

John Keogh of Hawthorn in Victoria, Australia, was awarded the technology prize for patenting the wheel, sharing the honour with the Australian patent office. David Schmidt of the University of Massachusetts won the physics prize for considering why shower curtains billow inwards. And Michigan-based televangelists Jack and Rexella Van Impe earned the astrophysics prize for showing that black holes fulfil all the technical requirements of hell. The Van Impes did not attend the ceremony, but Massachusetts Institute of Technology astrophysicist Walter Lewin, accepting temporary custody of their award, said the finding "will force us to rethink all of our ideas about black holes". ■

Illicit GM cotton sparks corporate fury

K. S. Jayaraman, New Delhi

Ten thousand hectares of unauthorized transgenic cotton have been found in India. The news has angered firms waiting to grow genetically modified (GM) crops in the country, and raises serious questions about the ability of developing nations to regulate the introduction of GM varieties.

The discovery — made by researchers from a company seeking to win legal authorization for the cultivation of a crop carrying the same gene — comes hard on the heels of the revelation that transgenic corn is growing in the wild in Mexico (see *Nature* **413**, 337; 2001).

Scientists at the Mumbai-based Maha-

ashtra Hybrid Seeds Company (Mahyco), in which US transgenic-crop giant Monsanto holds a 26% stake, discovered the cotton on farms in the western state of Gujarat, and informed the regulatory authorities. Monsanto holds a patent on the transgene involved, from the bacterium *Bacillus thuringiensis* (Bt). This encodes an insecticidal protein that allows the cotton to resist bollworm pests. Mahyco has been conducting field trials of the crop in India with a view to obtaining permission for its commercial use as early as next year.

The farmers in Gujarat had purchased the transgenic seeds from Navbharat, a company based in the state capital, Ahmedabad. Navbharat is thought to have developed the seed as a hybrid from transgenic seed imported from the United States.

The Genetic Engineering Approval Committee (GEAC) — a branch of the Indian Ministry of Environment and Forests, whose permission is required for any large-scale use of genetically modified organisms — has demanded an explanation from the company of how it came to be selling the seed without permission.

The discovery that the bollworm-resistant seeds are already available in Gujarat has infuriated Mahyco, which has spent US\$8 million on preparing to commercialize Bt cotton in India. Mahyco's managing director, Raju Barwale, says his company wants "strong and immediate

action" against Navbharat for its "blatant contravention of the legal and regulatory processes".

D. B. Desai, managing director of Navbharat, says his company will reserve comment on the allegations until he meets with GEAC officials. He declines to say where the seeds came from, although sources close to the company say they were brought from the United States (where Monsanto's GM seeds are freely available for sale) two-and-a-half years ago, and crossed with an Indian cotton variety to produce the hybrid. The sources add that Navbharat is likely to argue that it uses seeds from many sources, and lacks the technology to detect the transgenic strain in seed imported from the United States.

E. A. Siddiq, chairman of an Indian Department of Biotechnology committee that monitors transgenic crops, says: "This is a foretaste of a frightening situation where transgenics will be out of control and all over the place."

Arvind Kapur, vice-president of the All India Biotech Association, says that law-abiding biotechnology companies will suffer unless the government establishes tighter control of the sale of transgenic seeds.

To address the problem, argues P. K. Ghosh, a senior official at the biotechnology department, India needs to spend Rs100 million (US\$2 million) on six national laboratories equipped to monitor genetically modified organisms more closely. ■



Budding problem: thousands of hectares of transgenic cotton have been found in India.

US lays out bare bones of fossil protection package

Rex Dalton, Bozeman, Montana

A proposed package of new federal laws to protect palaeontological specimens on federal land and improve the management of palaeontology resources across the United States was announced last week.

The Paleontological Resources Preservation Act was introduced in the House of Representatives on 2 October. It has already been backed by the Society of Vertebrate Paleontology (SVP), which held its annual meeting in Bozeman, Montana, last week, and several federal agencies.

The act would increase the criminal penalties for stealing palaeontological specimens from federal land, unify federal policies for protecting and managing palaeontological resources, and better coordinate the recording of palaeontological discoveries.

"This bill will ensure that fossils will not be removed from the public domain, but are preserved for the enjoyment and education of all Americans for all time," says Richard

Stucky, president of the SVP and head of palaeontology at the Museum of Nature and Science in Denver, Colorado.

There are already laws against the unauthorized removal of palaeontological specimens. But the act would create a more uniform statute and raise the profile of protection efforts. It also requires that specimens are collected by qualified researchers with appropriate permits, who will then deposit the specimens in public institutions.

The act is modelled on the Archeological Resources Protection Act, enacted in 1979 to control looting and improve studies of Native American sites.

The new act was introduced by Representative James McGovern (Democrat, Massachusetts) and five other members, both Democrat and Republican. "I strongly believe we need a more comprehensive and thoughtful approach towards managing these public resources — an approach based on science not profit," says McGovern.

Palaeontologists hope the act will control the black market in looted palaeontological specimens. The illicit international fossil market was highlighted by a recent case in which dealers are charged with stealing a near-complete *Allosaurus* skeleton a decade ago from federal land in Utah. ■



Bangladeshis to sue over arsenic poisoning



Telling signs: drinking contaminated water causes calluses on the palms, leading to cancers.

Tom Clarke, London

The British Geological Survey (BGS) is preparing to defend itself against a threatened lawsuit alleging that it is partly to blame for what has been described as the worst mass poisoning in history.

The claimants are Bangladeshi villagers who drank arsenic-contaminated water from wells dug by the BGS during the 1980s and early 1990s. The number of claimants could run into thousands. They intend to file a group-action suit in Britain alleging that the BGS's failure to test the well-water for arsenic makes the agency responsible for their poisoning.

On legal advice, BGS officials declined to be interviewed about the allegations. But in a statement, the BGS vigorously denies culpability: "Any legal claims, which we regard as wholly misconceived, will be resisted."

The suit is being prepared by Leigh, Day and Co., a London-based law firm that specializes in this type of case. Some researchers fear that the threat of such suits may deter scientists from working on aid projects in the developing world.

The claim against the BGS is expected to hang on the accusation that, when its scientists returned to Bangladesh in 1992 to assess the safety of the water in the wells they had dug, they did not test for arsenic.

The BGS statement argues that, at this time, arsenic was not a recognized contami-

nant of groundwater in Bangladesh or regions with similar flood-plain geology: "There was no indication in the hydrochemical or hydrogeological literature by 1992 that arsenic might be associated with [river and delta plains]."

The arsenic crisis in Bangladesh was caused by the sinking of over one million wells by foreign aid agencies and the Bangladeshi government from the 1970s. The effort — intended to reduce the incidence of water-borne disease in the region — has instead put millions of Bangladeshis at risk of poisoning. "Bangladesh makes the Chernobyl disaster look like a Sunday-school picnic," says Richard Wilson, a Harvard University expert on arsenic poisoning.

The problem was not noticed until victims began showing external symptoms of arsenic poisoning: calluses on the palms and soles of feet, leading to skin cancers.

Most of the wells were sunk by UNICEF, the United Nations Children's Fund, in the 1970s and 1980s. The Bangladeshi government considered suing UNICEF, according to lawyers at Leigh, Day, but did not do so because there is a consensus that the arsenic problem couldn't have been anticipated then.

However, Bozena Michalowska of Leigh, Day claims that the BGS should have known about the potential for arsenic contamination in 1992.

Expert opinion on that point is divided. "Most hydrogeologists worth their salt would have known about it," claims John McArthur, a sedimentologist at University College London, who points out that reports of chronic arsenic poisoning started to emerge from neighbouring, and geologically similar, West Bengal in the mid-1980s. Dipankar Chakraborti, an epidemiologist at Jadavpur University in Calcutta, India, who identified the West Bengal cases, feels that the BGS should have been aware of his research. "They neglected the whole thing," he says.

But other experts point out that Chakraborti's work was not published in the international scientific literature, and only came to widespread attention in 1995. Paul Younger, a hydrogeologist at the University of Newcastle upon Tyne who has extensive experience of working in developing countries, says that "no one would have dreamt" of finding arsenic in river plains.

The fear of liability among foreign hydrogeologists, adds Younger, "could be a hell of a disincentive to performing low-cost schemes in developing countries that have been shown to be extremely effective".

Even Chakraborti fears that the case may deter others from working in Bangladesh and West Bengal. "If these agencies withdrew their support it would be a great loss," he says.

www.bgs.ac.uk

Spain's postdoc plan takes off

Xavier Bosch, Barcelona

The Spanish science ministry has announced the first 800 appointments to its new Ramón y Cajal programme, a tenure-track plan aimed at increasing competition for postdoctoral positions in the country's research institutions.

The Cajal programme, named after the Nobel-prizewinning neuroscientist, was launched earlier this year (see *Nature* 410, 1014; 2001). Successful applicants are offered five-year contracts, receive a salary of Ptas4.75 million (US\$26,300), which is similar to that of a university professor, and are able to run their own projects.

Of the new Cajal scholars, 169 will return home from positions abroad — something the programme was designed to encourage. Life sciences gains the most researchers, with 150 new appointments for genetics and molecular and cellular biology, followed by physical and space sciences with 80 and chemistry also with 80. But the ministry has complained of a lack of high-quality candidates in engineering and computer

science, with under 40 awards between them.

The programme will greatly increase the options open to postdocs in Spain — only 200 new contracts are available to them this year under other schemes. But the programme has been dogged by allegations that it has failed to break free of the favouritism that prevails in Spanish research institutions.

All researchers were required to obtain a letter of acceptance from the research centre



Science minister Birulés warned of problems.

they wished to go to before applying to the Cajal programme, and science minister Anna Birulés warned in June that some institutions were damaging the programme by giving such letters only to local candidates. Spanish Congress members have proposed that some of the remaining 1,200 Cajal posts should be granted without an acceptance letter.

www.mcyt.es/cajal

CERN management faulted over sudden budget crisis

David Adam, London

Soaring magnet costs, shaky project management and the lack of any contingency funds are being blamed by researchers for the cash crisis facing the Large Hadron Collider (LHC) currently being built at CERN, the European particle physics laboratory.

CERN says that the project, which is up to SFr850 million (US\$530 million) over its original SFr2.6-billion budget (see *Nature* **413**, 441; 2001), has been hit by large bills for equipment and civil engineering.

Officials at the laboratory, which straddles the border between France and Switzerland near Geneva, say they remain confident that the machine will collide its first protons as scheduled in 2006. But the source of the extra cash needed to meet this deadline remains unclear. CERN researchers and governors have asked why the project slipped so far into the red before the alarm was sounded, and there have already been calls for the resignation of CERN's director general, Luciano Maiani.



Luciano Maiani
faces criticism
from CERN staff.

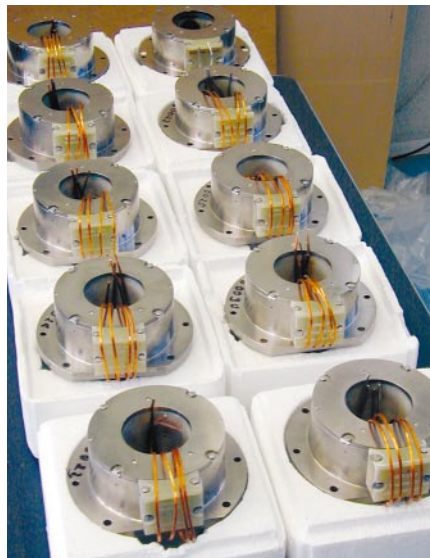
One CERN researcher, who declined to be named, says that poor financial supervision was a factor: "The management should have seen it coming, but financial controls were not done properly and accounting procedures were not careful enough."

Maiani argues that a thorough cost review was only possible after the closure of CERN's previous big accelerator, the Large Electron-Positron Collider, in November 2000. "Now is the time to act," he says. "A series of measures, including tight monitoring and control of spending, will be presented to our member states in November."

CERN officials plan to set up an 'office for spending control' to oversee its programmes. "This will enable the directorate to react in real time to any problem," says Maiani.

CERN spokesman Neil Calder says that the current financial troubles can be traced back to 1996, when the LHC was approved on a tight budget, with no contingency funds to pay for cost overruns. CERN was "not in a strong negotiating position" at the time, he says, explaining that it had to settle for what was on offer from member states, whose leaders were nervous about committing funds to a project similar in scope to the US Superconducting Supercollider, which had collapsed amid spiralling costs three years earlier.

Calder says that CERN originally intended to absorb part of any cost overrun by taking



Costly creations: prototype magnets for the Large Hadron Collider were unexpectedly expensive.

longer to build the machine or to pay off loans. "Our contingency was time," he says. But time is running out. Delays in getting the LHC operational will drain researchers and status from CERN. The rival Tevatron facility at Fermilab near Chicago, meanwhile, would enjoy an extended run at tracking down the Higgs boson, an elusive elementary particle that the LHC was largely designed to find.

Hopes that CERN's central operating budget would swallow some overspend have also proved optimistic. SFr150 million more than expected has already been spent on developing prototype magnets, but none of it has so far been drawn from the central budget. Insiders say that part of the problem in this case was a lack of in-house expertise, which meant that CERN had to pay for private contracts with companies that had little incentive to keep costs down. Quotes for the huge dipole magnets that will steer ions and particles through the LHC accelerator have also been significantly higher than expected.

Representatives from CERN's 20 member states are angry about the revelations — especially as CERN's management denied rumours of funding problems in June this year. When the problems were finally discussed by the CERN council on 20 September, delegates from the Netherlands called for Maiani's resignation, according to another delegate. Other council members, all of whom declined to be identified, think that he should stay to help sort out the mess. But they agree that Maiani and other senior CERN managers have some serious questions to answer. "There needs to be a change of climate at CERN," says one delegate. ■

Ireland weighs up options to buy into European research

Erica Klarreich, London

Ireland, which has boosted its research budgets in a bid to provide greater backing for innovation, is poised to join some of Europe's biggest science collaborations.

A report commissioned from an independent US group recommends that Ireland joins the European Molecular Biology Laboratory (EMBL) as soon as possible and that it then considers joining the European Synchrotron Radiation Facility (ESRF) and the European Southern Observatory (ESO). But it says that joining CERN, the European particle physics laboratory, would be too expensive.

The report was written by specialists in science and technology policy at the Georgia Institute of Technology in Atlanta, for Forfás, an advisory panel funded by the Irish government. It is likely to have a strong influence on the government's decision, expected in the next few months, on which projects to join, although officials say they are also accepting submissions from interested parties.

Ireland plans to spend 400 million euros (US\$368 million) a year on basic research, more than double its spending in the late 1990s. "In this changed funding environment, we decided it was time to have a serious look at joining some of these organizations," says John Fallon, a senior official in the Department of Enterprise, Trade and Employment, which is responsible for science policy.

The report, published in its final form on 15 September, says that the EMBL's aims fit well with Ireland's research priorities, and notes that membership would be cheapest of the facilities considered, at 450,000 euros per year.

The growing importance of synchrotron research and the public appeal of astronomy led the report's authors to recommend the ESRF and the ESO. In advising against joining CERN, the report cites a lack of Irish high-energy physicists, together with the high membership cost of between 6.5 million and 7.5 million euros per year.

But Irish physicists and some government officials disagree, arguing that membership of CERN would encourage students to enter the field. "It's a chicken-and-egg situation," says Roger O'Connor, director of business and technology in the Department of Public Enterprise. "Particle physics will look much more attractive to researchers looking for doctoral subjects if Ireland becomes a member of CERN." ■

US universities body backs tighter rules on conflict of interests

Washington The Association of American Universities is recommending a tightening of existing federal guidelines on conflicts of interest in research.

In a report released this week, the association says that individual researchers should disclose all relevant financial interests and sources of research funding. It adds that scientists who are engaged in research involving human subjects should avoid having any financial stake in their work.

The report also says that institutional policies need to be strengthened, and that universities and research centres should develop their own policies regarding disclosure of institutional financial interests, including equity holdings, royalty agreements and the financial holdings of senior university officers.

♦ www.aau.edu

No contest over new Max Planck Society head



Peter Gruss: backs stem-cell research.

Munich Developmental biologist Peter Gruss looks set to become the next president of Germany's Max Planck Society. A Max Planck committee will vote on the issue next month, but Gruss is the only candidate on the shortlist.

Gruss, currently a director at the Max Planck Institute for Biophysical Chemistry in Göttingen, is best

known for identifying the central role of the *Pax6* gene in the development of the eye and forebrain in mice. His group is now searching for other genes involved in development of the nervous system.

Gruss recently became involved in the highly contentious stem-cell debate when he backed importing human embryonic stem cells from other countries. Isolation of the cells from human embryos is currently forbidden in Germany.

US websites closed in anti-terrorist bid

San Francisco Researchers using certain US government websites and online databases are preparing for a spate of 'Page not found' messages.

Several agencies, including the Centers for Disease Control and Prevention (CDC) and

the Environmental Protection Agency (EPA), have shut down sites they say could contain information useful to terrorists.

The EPA's National Exposure Research Laboratory website, for example, is closed pending a review of its contents. The site, which is used by toxicologists, includes information on the risks posed to people and the environment by spills from chemical plants. Pages pulled from other sites include the National Imagery and Mapping Agency's water-pipeline maps and a CDC report on chemical terrorism.

Several Internet search engines store web pages, however, and much of the information is still available from these sites.

European satellites to spot gorillas in the mist

Paris Sites of key cultural and scientific value will be monitored from space under a new collaboration between the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the European Space Agency.

UNESCO's World Heritage Committee currently lists 690 such sites. One of these, the central and east African forests that are home to threatened gorilla populations, will be used in a pilot study to select the most appropriate monitoring technology. Earth-observation satellites are likely to play a key role. The collaboration was announced at a meeting of the International Astronautical Congress held in Toulouse, France, last week.

♦ www.unesco.org/whc

Volunteers flock to donate computer time

London Two research projects are celebrating milestones in the use of distributed computing, in which volunteers donate their computers' idle time to analyse scientific data.

A drug-screening project run by the University of Oxford and the National Foundation for Cancer Research, based in Bethesda, Maryland, has attracted one million participants since its launch this April. The scheme's screen-saver software has tested 3.5 billion different molecules to determine whether they attach to two proteins implicated in cancer development.

SETI@home, which examines radio-telescope data for signals from extra-terrestrial life, has so many participants that it cannot collect data fast enough. The researchers behind the scheme, at the University of California, Berkeley, are now planning to widen the area of the electromagnetic spectrum from which they gather signals in order to keep up with the demand from volunteers.

♦ members.ud.com/vypc/cancer

♦ setiathome.ssl.berkeley.edu

Privacy plan could thwart animal-rights protesters

London Directors of companies who have faced violent protests from animal-rights activists may soon be able to keep their home addresses private, thanks to a proposed change in British law.

The draft regulations, published last week, were prompted by violence and intimidation directed against staff and directors of Huntingdon Life Sciences, a drug-testing firm based in Cambridgeshire. Huntingdon's managing director Brian Cass was attacked outside his home in February by three masked assailants wielding baseball bats.

British company directors are currently required to make their home addresses publicly available. The regulations will allow directors to transfer this information to a private register if they can prove that they, or someone who lives with them, are at risk from violence or intimidation.

Pandas endangered by high blood pressure

Tokyo Battling extinction can be stressful, and China's giant pandas appear to be showing the strain. Some are suffering from high blood pressure, say researchers.

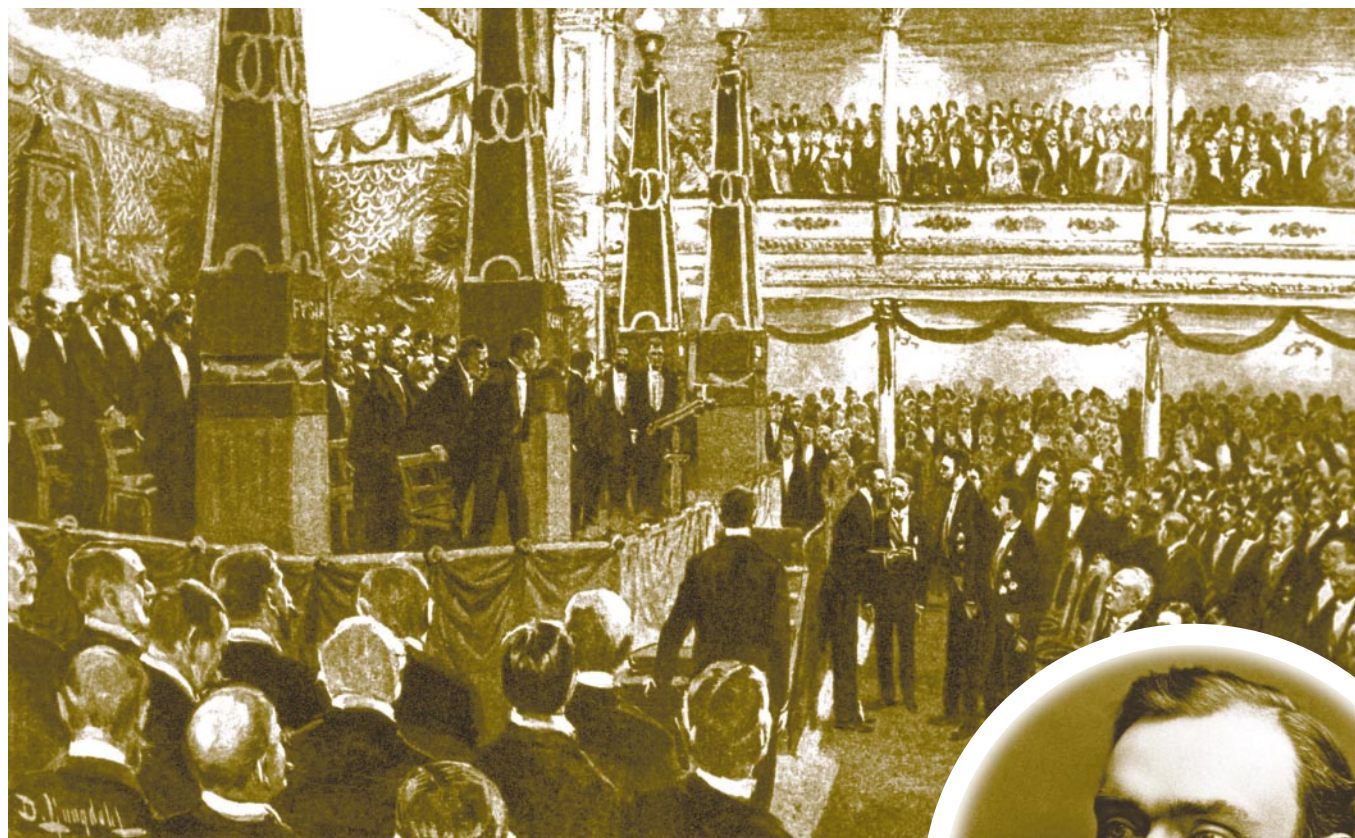
Chen Yucun, director of the Fuzhou Giant Panda Research Centre in China's eastern Fujian province, says that high blood pressure could account for many of the ailments suffered by pandas, and that they should have regular checks.

The centre recently treated a 21-year-old panda with hypertension, and succeeded in bringing the animal's blood pressure down to normal levels by administering double doses of the drugs used to treat high blood pressure in humans.

The centre claims to have achieved several breakthroughs related to the breeding of giant pandas, including the cloning of panda embryos and the world record for the oldest panda to give birth.



Stressed out: raised blood pressure is a headache for China's pandas, say researchers.

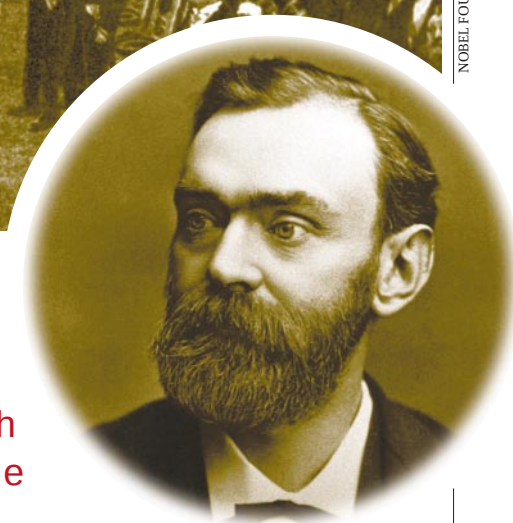


NOBEL FOUNDATION

NOBEL FOUNDATION

Eyes on the prize

The Nobels mark their centenary this week. Their prestige is unquestioned, but does the way in which winners are selected reflect the way science is done in the twenty-first century? Trisha Gura investigates.



For a few days each October, scientists who were previously famous only within their narrow fields of endeavour are fêted as heroic pioneers who have extended the boundaries of human knowledge. Areas of research that might normally struggle to gain media attention find themselves given top billing.

Such is the prestige of the three science Nobel prizes. The announcement of this year's winners (see pages 553–554) marks the awards' centenary — a good time to reflect on how they came to acquire their present status, and to ask whether the procedure for choosing the winners is in tune with twenty-first-century science.

The Nobels perform an impressive public-relations job for research — or at least for those disciplines that the prizes honour. They create an inspiring image of lone pioneers receiving the ultimate accolade for their individual brilliance. But now that science has evolved into a more collaborative process in which projects can command billion-dollar

price tags and involve thousands of scientists, is this image really appropriate? And how do the few key individuals to be rewarded get selected? The details remain mysterious, as the Nobel committees' deliberations are kept secret for 50 years after each prize is awarded. But as historians of science trawl through earlier archives, they have found that the selection process has, on occasion, been eccentric.

Winning lines

When he wrote his will, Alfred Nobel probably had no idea what it would lead to. Fabulously wealthy from his invention of dynamite, Nobel bequeathed the majority of his estate to create five prizes to honour achievements that have “conferred the greatest benefit on mankind” in chemistry, physics, physiology or medicine, literature and the promotion of peace.

Nobel charged four Scandinavian institutions with selecting the winners: the Royal Swedish Academy of Sciences, responsible for physics and chemistry; the Swedish Acad-

emy, for literature; the Karolinska Institute, Sweden's leading medical school, for physiology or medicine; and a committee appointed by the Norwegian parliament, for peace. (The economics prize is a modern invention, established by the Bank of Sweden and first awarded in 1969.)

The prizes' value was unprecedented. When they were first awarded in 1901, five years after Nobel's death, the value of each represented about 30 times the salary of a university professor. Today, a lone winner of a Nobel will take home 10 million Swedish Kronor, or about \$940,000.

The sums involved certainly helped to build the Nobel legend, but money cannot completely explain the prizes' unique status. A handful of other science awards — some of which were set up to reward disciplines not covered by the Nobels — now offer handsome payments (see ‘Not quite a Nobel...’, page 564). Yet none have quite the same cachet — how many winners of the Crafoord prize can you name?



Past and presentations: the first Nobel award ceremony in 1901 (left) and last year's version (above) — prestigious legacies of Alfred Nobel (left, inset).

Science historians argue that the Nobels owe part of their prestige to the geopolitical climate into which they were launched — and the calibre of some of the early winners. Nobel willed that the prizes should be truly international, which gelled with the intense, but peaceful, national competition that characterized the end of the nineteenth century — the modern Olympics had just been launched, and world fairs were in vogue. Within the first five years, Nobel laureates included such luminaries as Wilhelm Conrad Röntgen, the discoverer of X-rays, physiologist Ivan Pavlov, and Pierre and Marie Curie, rewarded with Henri Becquerel for their work on radioactivity.

Today, the stakes are even higher. In addition

to the monetary value of the prize, laureates can expect their labs to receive an immediate boost in funding from public and private sources, and unprecedented opportunities for travel. For those who enjoy the celebrity, there is also the prospect of book contracts and television appearances.

"In the first few years after you get the prize, granting is probably biased in your favour," says Hamilton Smith of Celera Genomics in Rockville, Maryland, who shared the physiology or medicine prize in 1978 for the discovery of DNA-cutting restriction enzymes.

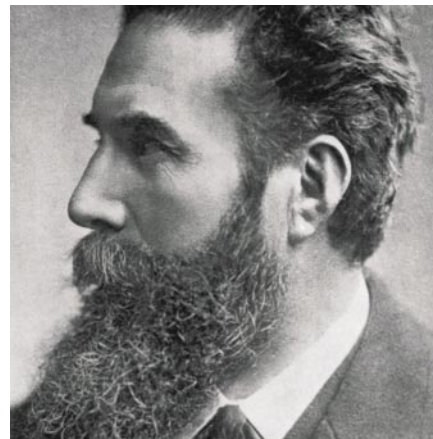
Arvid Carlsson of Gothenburg University in Sweden, who shared last year's medicine or physiology prize for his work on the

neurotransmitter dopamine, says that obstacles which had been frustrating his efforts to launch a spin-off company to exploit a drug for Parkinson's disease, including financing and accommodation on the university campus, suddenly melted away. "A number of problems were completely eliminated," he says.

Fame game

Nobel laureates can also become public figures, whose views are sought by those in power. Take Peter Doherty, who shared the prize in physiology or medicine in 1996 for his work on cell-mediated immunity. Doherty, who works at St Jude Children's Research Hospital in Memphis, Tennessee,

SPL: CORBIS; NOVOSTI/SPL

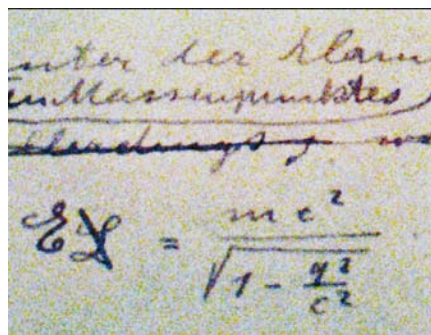


Shining stars: luminaries such as Pierre and Marie Curie (left), Wilhelm Röntgen (centre) and Ivan Pavlov (bearded, right) helped to cement the Nobel legend.

was the first Australian to win a Nobel since 1975. "I have been very much on the public stage in Australia," says Doherty, who has found himself projected as a spokesman for science, and given access to the country's leading politicians. He jokes about acquiring a celebrity akin to that of "a sort of minor figure in a coffee commercial".

No wonder, then, that scientists worldwide are fascinated by the process for selecting winners. This had to be painstakingly worked out after Nobel's death, as he left no clear instructions. The inventor had not consulted with any of the institutions he named, nor did he illuminate his two executors — his personal assistant Ragnar Sohlman and engineer Rudolf Lilljequist, who had no prior association with Nobel.

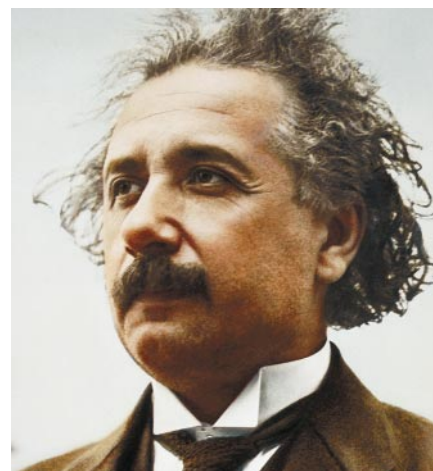
Rows erupted when the executors, academic institutions, the Swedish government and Nobel's family — who received a tiny fraction of his inheritance — sat down to thrash out a legal statute. For the science prizes, the sparring factions eventually settled on a formula that remains largely unaltered a century later. Nobel's wish that the prize be awarded for work conducted in the preceding year was immediately thrown



All in the mind: Einstein's general theory of relativity was dismissed as mere speculation by the Nobel selection committee.

out — it simply did not give enough time to assess the impact of a discovery. And a rule that the prize could be split between two discoveries if both were thought to merit an award was inserted.

To be eligible for consideration, a candidate must be proposed by a nominator from one of six categories: members of the awarding institutions; members of a Nobel committee selected for each prize; previous winners; professors of physics, chemistry and medicine at Swedish and other Nordic



AP/CORBIS

universities and technical colleges that existed in 1900; chairs in these fields at six other universities or faculties of medicine selected by the awarding institutions; plus scientists invited to participate on an ad hoc basis.

The Nobel committees, made up of between three and five people, lie at the centre of the decision-making process. There are no limits on members' nationalities, but historically most have been Swedish.

Japan seeks a record haul

In December 2000, as part of its Science and Technology Basic Plan, the Japanese government set its scientists an ambitious goal: 30 Nobel prizes in the next 50 years. This would mean a fivefold increase in the success rate experienced over the past half-century.

Few Japanese researchers take this statement seriously. As one researcher at the University of Tokyo's Graduate School of Medicine says: "Could the government of any other country get away with making such ridiculous promises?" How a national drive for greater Nobel success should be conducted remains unclear, and there is the potential for efforts to backfire. An information office established in April at the Karolinska Institute in Stockholm by the Japan Society for the Promotion of Science has already been dismissed by some scientists as a veiled — and vain — attempt to lobby the medicine Nobel committee.

But the plan has at least sparked a debate on why Japanese scientists are so poorly represented in the ranks of Nobel laureates. Some Japanese scientists argue that the problem is a fundamental lack of creativity, stemming from an educational and social system that promotes rote learning and conformity. But James Bartholomew, a historian of Japanese science at Ohio State University in Columbus, rejects this self-deprecating stereotype. "The image of the Japanese as unoriginal is unfounded," he says.

Nevertheless, many researchers believe that Japan's academic 'koza' system, which places immense power in the hands of university professors and limits the freedom of younger investigators to set their own research agendas,

has a stifling effect. "Everyone in our lab had to do projects dictated by our professor," says one molecular biologist, now in the United States. "But we each had our own 'shadow projects', which we could only work on when he was not around."

Hideki Shirakawa of the University of Tsukuba, who shared last year's chemistry

Kenichi Fukui (back row, centre) with the other scientific laureates of 1981.



Hideki Shirakawa receiving his chemistry Nobel last year.

Nobel for his work on electrically conducting organic polymers, argues that improving the situation may require changes to the laws that govern Japan's public university system. "Younger scientists

have a very restricted range at universities," he says. "They have no chance to change supervisors unless they go abroad."

Language may also be a barrier — researchers who are not at ease with English inevitably struggle to gain international recognition. Bartholomew points to Kenichi Fukui of Kyoto University, who shared the 1981 chemistry Nobel for his theoretical insights into the mechanics of chemical reactions. Fukui, who died in 1998, was criticized for his "insufficiently elegant" English formulation of his ideas even after he won, says Bartholomew. How many more deserving candidates failed to be nominated because of their poor English skills remains unclear. Lurking in the background, though seldom voiced directly, are fears that Japanese researchers are discriminated against by the European and North American scientific establishment.

But some members of that establishment argue that Japanese scientists must shoulder some of the blame. Speaking on a visit to Japan in July, Anita Aperia of the Karolinska Institute, a member of the medicine Nobel committee, questioned the approach of some of the nominations coming from Japan. "People have to realize that the nominations are for discoveries, not for lifetime achievements," she said.

Whatever the reasons for Japan's poor showing in the Nobels, few expect an official target to make an enormous difference. "We can't be short-sighted," says Reiko Kuroda, a biophysicist at the University of Tokyo and a member of the Council for Science and Technology Policy, the government's premier science advisory committee. "Nobels are an outcome of good research culture, not a target to be directly aimed at."

David Cyranoski, Tokyo



AP

AP



Into the limelight: both Arvid Carlsson (left) and Peter Doherty say their Nobels altered their lives.



I have very much been on the public stage ... like a sort of minor figure in a coffee commercial. Peter Doherty

Members are selected by the awarding institutions, and their terms are limited to nine years. Many of the conventions that determine the committees' workings are unwritten. "It takes a long time to learn the rules," says Cecilia Jarlskog of CERN, the European Laboratory for Particle Physics near Geneva, who headed the physics committee until 1999.

Usually, the committees' word is final. But they are monitored by the awarding institutions, which on occasion have pulled rank. In 1906, for instance, the Royal Swedish Academy of Sciences rejected the chemistry committee's majority decision to award the prize to Dmitri Mendeleev for his periodic table of the elements. This intervention was a result of the powerful influence exerted by Swedish chemist Svante Arrhenius, who won the chemistry prize in 1903 for his theory of electrolytic dissociation — of which Mendeleev had been a prominent critic. Mendeleev died the next year and never won a Nobel.

Picking winners

Today, as in the early 1900s, invitations for nominations are sent out in September with a deadline of the following February. Committee members spend their summers reading through the stacks of evaluations and publications.

For three-quarters of a century, this was all the public knew about the selection process. But in 1975, the Nobel Foundation decided to open the archives on each prize after 50 years. Science historians have since pored over details of the nominations, and of the views expressed by members of the Nobel committees, for the early prizes. If the archives are any indication, the statutes generated confusion and contention from the start. Nominators were not sure how to interpret wording such as "greatest benefit to mankind" or even how to define physics or chemistry.

For example, in 1923, the solar astrophysicists George Ellery Hale and Henri Deslandres had emerged from the Nobel physics committee's earlier deliberations as

leading candidates. But according to historian Robert Marc Friedman of the University of Oslo, author of *The Politics of Excellence: Behind the Nobel Prize in Science* (Henry Holt, New York, 2001), newer members of the committee decided that astrophysics was a subdiscipline of astronomy, rather than physics, and so was not eligible.

Friedman says that many of the early committees were rife with strategizing and personal biases. "The distance between the high ideals declared and the actual practices are even more apparent than I had originally anticipated," he says.

One classic struggle surfaced in the delay in awarding prizes to the physicists Max Planck and Albert Einstein. According to Friedman and other historians, this was mainly because the Swedish scientific community in the early 1900s embraced experi-



Difficult choices: Cecilia Jarlskog believes that selection processes may have to change.

mental physics and dismissed theory as mere speculation. As late as 1921, Bernhard Hasselberg of the Royal Swedish Academy of Sciences, a member of the Nobel physics committee, intervened from his sickbed to protest against Einstein being rewarded for his work on general relativity. "It is highly improbable that Nobel considered speculations such as these to be the object of his prizes," Hasselberg wrote.

Faced with impasse, the Royal Swedish Academy of Sciences deferred the award of the 1921 physics prize. A solution was found the following year by committee member Carl Wilhelm Oseen of Uppsala University, who suggested that Einstein be awarded the reserved 1921 prize for his discovery of the law of the photoelectric effect. But Oseen was careful to delineate the empirically proven law from the quantum theory that lay behind it — the idea that light can behave as particles as well as waves.

Rumour mill

It is not clear to anyone outside the committees or awarding institutions whether such internal battles still take place, as the deliberations remain shrouded in secrecy. In this information vacuum, rumours abound about candidates and their associates or host institutions lobbying for Nobel recognition. "I hear there are a lot of scientists and people around them who are in touch with the Karolinska outside the regular nomination procedures," says Carlsson. The decision by the Japan Society for the Promotion of Science to establish an information office at the Karolinska Institute, for instance, has been viewed in this light (see 'Japan seeks a record haul', opposite).

Many observers argue that overt lobbying is unlikely to be effective — and is probably counterproductive. The committees "are not exactly naive," says Doherty. Jarlskog stresses that flooding the Nobel committees with nominations for a particular individual will not work, as winners are not chosen by tallying the nominations.

Some scientists argue that the entire process should be opened up to public scrutiny. "I think this is completely out of phase with the current face of science," says one critic, who asked not to be identified. "If the prize is going to have the capacity to rewrite history, which it does, the system ought to bear some accountability." But others argue that a more transparent process would only unleash torrents of discontent. "I would not like to see candidates know who nominated them," adds Carlsson.

Jarlskog argues that truly deserving candidates tend to be nominated year after year, which means that their claims are eventually recognized. "We assume that it is better to be late than to be wrong," she notes. But thanks to rules that prevent posthumous awards, even if the candidate was alive when

► nominated, the wait proves too long for some candidates. "There is a lot of luck in this," Carlsson says. "In my case, I had to see to it that I lived sufficiently long."

Three-way splits

But the most vociferous debate surrounds the 'three-person' rule. This is a 1968 amendment to the statutes, covering all Nobel prizes, which states: "in no case may a prize be divided between more than three persons".

The rule sparked intense controversy three years ago, when many scientists were outraged at the exclusion of Salvador Moncada, director of the Wolfson Institute for Biomedical Research at University College London, from the medicine prize (see *Nature* **395**, 625–626; 1998). That prize, for research on signalling by nitric oxide in the cardiovascular system, was shared by Ferid Murad of the University of Texas Medical School in Houston, Robert Furchgott of the State University of New York and Louis Ignarro of the University of California, Los Angeles.

Although no one doubted the winners' contributions, previous laureates led by César Milstein of the Laboratory of Molecular Biology in Cambridge, winner in 1984 for his discovery of monoclonal antibodies, argued that Moncada had been an equally pivotal figure. Moncada today declines to discuss his bitter disappointment. But Murad argues that there were "one or two others in line" before Moncada, and says he is disappointed by the negative publicity gen-

Not quite a Nobel...

The Nobel prize may be the ultimate scientific honour, but over the years various other awards have been established that carry immense prestige — and, in some cases, rich monetary rewards.

Recognizing that Alfred Nobel's will left many scientific disciplines out in the cold, the Royal Swedish Academy of Sciences in 1980 established the Crafoord prize (www.kva.se/eng/pg/prizes/crafoord/index.asp). Awarded each January, this US\$500,000 prize recognizes outstanding researchers in fields including mathematics, astronomy, Earth sciences and ecology.

The four Swiss–Italian Balzan prizes (www.balzan.it), worth US\$620,000 each, are similarly broad in their scope, and since 1961 have been awarded each year for outstanding achievements in a broad range of intellectual endeavours including physical, mathematical and natural sciences, and medicine. The US\$100,000 Israeli Wolf prizes (www.aquanet.co.il/wolf/wolfpriz.html) similarly cover a range of areas including mathematics, agriculture, chemistry, physics and medicine.

The discipline most obviously passed over by Nobel, mathematics, has high-profile awards of its own. For the field's young guns, aged 40 years or less, up to four Canadian Fields medals (elb.zib.de/IMU/medals/index.html) are awarded every four years. They are worth only US\$9,500, but their value in terms of career enhancement is huge. And the Norwegian government this year announced that it is establishing a US\$570,000 Abel prize (see *Nature* **413**, 100; 2001), without age restriction, intended as a direct counterpart to the Nobels. The first prize will be awarded in 2003.

But not all of the other top prizes are there to fill in the gaps left by Nobel's will. The 54-year-old Albert Lasker medical research awards (www.laskerfoundation.org/awards/awards.html) — often dubbed the 'American Nobels' — are presented each year in September. There are three awards, for basic medical research, clinical medical research and public service, each worth \$50,000, and they are a good indicator of further success. More than half of the Lasker winners since 1962 have gone on to receive a Nobel. T.G.

erated by the protests. "It was a mess and totally inappropriate," he says.

This particular controversy might never be entirely settled, but it highlights the three-person rule's limitations. In many fields, it is increasingly difficult to identify three or fewer individuals responsible for a particular discovery. In disciplines such as high-energy physics and genomics, it is practically impossible. "With collaborations getting bigger and bigger, it is not obvious who the driving force was," says Roger Cashmore, director of research for collider programmes at CERN.

Celebrity shared

Cashmore suggests that, in such cases, the prize might be awarded to an institution, as has been done for the peace prize on several occasions, most recently to the medical aid charity Médecins Sans Frontières in 1999. "It does not make as huge an impact, but it does award the institution enormous recognition and kudos," he argues. But so far, the awarding institutions for the science prizes have declined to allow institutional awards.

Although such awards might ease the particular headaches created by disciplines such as high-energy physics, they would leave the general issue of the three-person rule's potential for arbitrariness unresolved. "We ought to create something that gets to the real heart of what science is supposed to be about," advocates Friedman.

But for every voice in favour of relaxing the three-person rule, there is another arguing for tradition. Says Murad: "If you change it from three, should it be given to four, ten, twenty? Where do you stop?"

Others argue that relaxing the rule would destroy the mystique that enables Nobel laureates to become influential spokes-

people for science. "It's the star system," says Doherty. "By giving the prize to hundreds of people, you will rapidly lose that effect."

"One should be very careful," agrees Jarlskog. But she believes that the awarding institutions may revisit the statutes to consider whether adjustments are needed to reflect the changing face of science. "They have to, and I think they eventually will," she says.

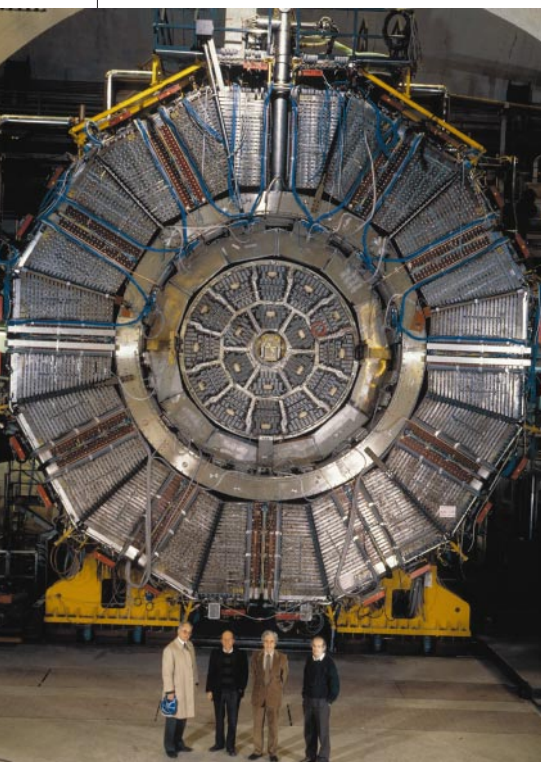
Meanwhile, some Nobel laureates suggest that one way to avoid the arguments would be to do away with the prizes altogether. "Einstein would be Einstein without the prize," says Jack Steinberger of CERN, who shared the 1988 physics Nobel for working out how to create intense, high-energy neutrino beams, and for discovering the muon neutrino.

Maybe so, but that begs the question: would Jack Steinberger still be Jack Steinberger, Peter Doherty be Peter Doherty, or Hamilton Smith be Hamilton Smith? They are all brilliant scientists, but their reputations were greatly enhanced by their Nobel awards. Einstein, on the other hand, was a one-off — a uniquely charismatic genius whose iconic status did not depend on the approval of the Nobel physics committee.

For those scientists who can only aspire to more everyday levels of excellence, the Nobels represent an unparalleled opportunity to propel the research that they love onto the public stage, and perhaps to use their new-found celebrity status for the good of science. For these and other reasons, it is a safe bet that each October for decades to come, many of the world's leading scientists will continue to dream of receiving that life-changing phonecall from Stockholm. ■

Trisha Gura is a science writer in Cleveland, Ohio.

► www.nobel.se



High-energy physics involves collaborations, which makes it hard to pick individual winners.

Public trust requires disclosure of potential conflicts of interest

Evidence that funding relationships can distort objectivity is growing — but money isn't the only source of bias.

Sir — We commend your adoption of a policy that requires disclosure of financial interests and biases of authors who publish in *Nature* (Opinion, *Nature* **412**, 751; 2001). Your effort to achieve transparency is an important response to the growing evidence of the adverse impacts that funding relationships can have on the integrity of science generally and on the reputations of individual institutions and researchers specifically.

More stringent conflict-of-interest policies are being adopted by universities, scientific societies and journals in the light of recent examples of publication and editorial interference by research funders; scientific misconduct; and even deaths of human testing volunteers. Several of the US medical schools receiving the most National Institutes of Health funding have been developing principles for conflict-of-interest policies (*Nature* **408**, 630; 2000),

recommending full public disclosure of research funding when research results are presented or published.

For university policies to be effective, science journals must also adopt and enforce policies of full public disclosure of funding sources and other potential financial conflicts. Several medical journals have announced plans to adopt more stringent, uniform conflict-of-interest policies in response to increased collaboration between industry and universities.

We support *Nature's* leadership role in this extremely important issue, and challenge other scientific journals to adopt and enforce similar policies in a global effort to maintain public trust in the integrity of science.

Steven Gurney, Jennifer Sass

Health and Environment Program, Natural Resources Defense Council, 1200 New York Avenue NW, Suite 400, Washington, DC, 20005, USA

Declaring interests in the fight for good science

Sir — Scientists for Global Responsibility (SGR) welcomes with enthusiasm your new policy of asking authors to declare their financial interests (Opinion, *Nature* **412**, 751; 2001). We congratulate *Nature* for taking a bold position on an issue of great importance.

SGR is particularly concerned that financial support from certain sources can skew the aims of research and distort the objectivity that is fundamental to good science. We have recently published a booklet entitled *An Ethical Career in Science and Technology?* (see <http://www.sgr.org.uk/ethics.html>). We hope that authors everywhere will support this policy and that scientists, their employers and those publishing their work will share the responsibility for ensuring an ethical approach to conducting and reporting scientific research.

One potential risk with a policy of voluntary disclosure is that a large proportion of authors may opt not to disclose. Yet mandatory disclosure risks penalising scientists who are contractually required to withhold information about their interests. *Nature's* approach may well be the best in the absence of a clear

solution to this issue, but in the long term all science funding should be open to public scrutiny.

Vanessa Spedding

Scientists for Global Responsibility, PO Box 473, Folkestone CT20 1GS, UK

Financial interests are not the only bias factor

Sir — Personal financial interests (Opinion, *Nature* **412**, 751; 2001) are relatively minor among the many factors relevant to potential bias. To think that bias will be meaningfully corrected by requesting disclosure of personal financial interests is naive or even harmful, as it could give a false assurance that bias has been checked.

No journal, *Nature* included, could ever guarantee the veracity of everything it publishes. In my experience, peer review simply screens gross and inappropriate statements; final or near-final answers come solely from the iteration of experiments in different hands. That is the only process of conditional verification known to science.

Nature has joined other special-interest publications in giving delusory guarantees of veracity, rather than continually warning and helping readers to develop their own critical skills. A more mundane, but altogether more admirable,

standard of publishing sincerity would be to print a disclaimer about implied standards of truthfulness on each issue's contents page.

Gio Batta Gori

The Health Policy Center, 6704 Barr Road, Bethesda, Maryland 20816, USA

Exciting future planned for birthplace of genetics

Sir — Fabio Salamancina in his Correspondence "Keeping Mendel in mind" (*Nature* **412**, 118; 2001), responding to the News story "Museum suffers spiritual cramps over Mendel's work" (*Nature* **410**, 6; 2001), expresses concern that the Mendel museum in the monastery in Brno is under threat. We would like to inform readers of an initiative to reinstate the monastery as a place of scientific discovery.

The proposal is that the monastery site could eventually house a research institute, a conference centre with a modern lecture hall and a museum of genetics, as well as providing courses for graduates and schoolchildren alike. It could create a forum for discussions on genetics and the wider ethical issues. Abbot Evzen Martinec has agreed that much of the site could be adapted to this end, if the funds could be found, and the city of Brno is keen to use the redevelopment of the site as a focus for urban renewal. President Vaclav Havel has given the project his full support.

Plans for the first exhibition in 2002 are going ahead, initiated by Old Brno monastery and curated by Marina Wallace and Martin Kemp in collaboration with the Mendelianum, Brno. An inaugural conference, "Genetics after the Genome", will be held concurrently.

Although seed money is available for a pre-feasibility study to be overseen by the architect Eva Jiricna, more money is needed for the project. Further information can be obtained from anna@nt.imp.univie.ac.at.

Kim Nasmyth*, Dieter Schweizer†

**Research Institute of Molecular Pathology, Dr Bohrgasse 7, A-1030 Vienna, Austria*

†Institut für Botanik der Universität Wien, Abteilung für Zellbiologie und Genetik, Rennweg 14, A-1030 Vienna, Austria

Debate over language's link with intelligence

Sir — M. Piattelli-Palmarini, in his review (*Nature* **411**, 887–888; 2001) of the book *Pathways to Language*, criticises authors Kyra Karmiloff and Annette Karmiloff-Smith for misrepresenting the linguistic

and cognitive abilities of children with Williams syndrome (WMS). Contrary to the book's authors, your reviewer claims that "children with Williams syndrome have a barely measurable general intelligence" but "an exquisite mastery of syntax and vocabulary", although they are "unable to understand even the most immediate implications of their admirably constructed sentences".

Based on our own extensive research on WMS in English and Italian, we disagree with your reviewer. First, intelligence is certainly testable in WMS. In older children, the average performance IQ is around 60, and many score much higher. By the age of 8–10 years, children with WMS are typically functioning at a mental age of 5–6, an age at which normally developing children display sophisticated vocabulary and complex grammar.

Second, syntax is far from perfect in WMS. In younger children, lexical and grammatical development are delayed; in older children, difficulties and errors persist. Overall levels of syntax never exceed mental age.

Third, comprehension abilities in WMS are often puzzling, but they are much more sophisticated than Piattelli-Palmarini implies. Individuals with WMS are extremely interesting for research on language, cognition and social functions, because they display an unusual profile of strengths and weaknesses that may be linked to the genetic alterations responsible for this syndrome. But they are not language savants, and do not provide evidence for intact language in the absence of measurable intelligence.

In short, contrary to your reviewer, we believe that Karmiloff and Karmiloff-Smith 'got it right'.

Elizabeth Bates*, Helen Tager-Flusberg†, Stefano Vicari‡, Virginia Volterra§

*Center for Research in Language 0526, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA

†Laboratory of Developmental Cognitive Neuroscience, Department of Anatomy and Neurobiology, Boston University School of Medicine, 715 Albany Street, Boston, Massachusetts 02118-2526, USA

‡Servizio di Neurologia e Riabilitazione, IRCCS, Ospedale Pediatrico Bambino Gesù, Lungomare Guglielmo Marconi 36, I-00058, Santa Marinella, Rome, Italy

§Istituto di Psicologia, CNR, Viale Marx 15, 00137 Rome, Italy

Concerns highlight need to make faster decisions

Sir — Your News story "Venture capital concerns academics" (*Nature* **413**, 95; 2001), on the involvement of the

University of California, San Francisco (UCSF) with the Burrill venture fund, suggests a plan hatched by investors and university officials to exploit our faculty without our consent.

The conflicts for faculty and students resulting from university–industry collaborations have been well documented in, for example, The Business–Higher Education Forum's *Working Together, Creating Knowledge: The University–Industry Research Collaboration Initiative* (American Council on Education, Washington; 2001. <http://www.acenet.edu/bookstore/index.cfm?pubID=230>). It would be a pity if articles such as your News story derailed the recent improvements in relations between universities and industries by artificially setting academic scientists against those marketing their discoveries.

The lofty 'bench-to-bedside' goal we aspire to requires that scientists facilitate the transfer of their ideas into the commercial world. At UCSF we scientists, of our own volition, submit about 150 disclosures a year to our Office of Technology. Far too few of these disclosures result in useful products. A committee I chaired identified several contributing factors, including lack of funding for proof-of-principle research and for intellectual-property costs, lack of business acumen among our faculty, and cumbersome procedures. To speed up the process, we developed a non-restrictive agreement with Burrill that we hope will unclog the pipeline with no loss of faculty autonomy or of Burrill's autonomy to set up similar agreements with others.

The agreement introduced no significant changes in the way UCSF operates and required no faculty input. Nonetheless, as a public institution we are especially sensitive to public perception, and so our proposal was offered for comment to several advisory groups, including the Academic Senate, the official voice of the UCSF faculty. Some members of the senate felt that two months was not sufficient time to identify any potential conflicts of interest. It was their dissatisfaction that was reported in your News story. (And as this correspondence goes to press, the proposal seems increasingly unlikely to come to fruition).

You have nicely illustrated a real conflict of interest in industry–academic interactions. Industry needs a hierarchical structure that allows rapid decisions. Universities have a diffuse decision-making structure that can be slowed to glacial speeds because of lack of faculty time and diversity of opinion. If university research is truly to benefit society we need to find ways to accommodate faculty decision-making processes to the honest

needs of our industrial counterparts.

Regis Kelly

Executive Vice Chancellor, UCSF, Department of Biochemistry & Biophysics, University of California, San Francisco, California 94143, USA

Sound basis for research

Sir — Your News story "Fears for basic science as Bush backs use of investment criteria" (*Nature* **413**, 5; 2001) expresses concerns over the fate of research funding and policy under the Bush administration, owing to the establishment of research-and-development performance criteria. The Bush administration is indeed examining explicit criteria for federal investment in research according to its *President's Management Agenda*.

This initiative was prompted by two excellent reports by the Committee on Science, Engineering, and Public Policy (COSEPUP) — *Evaluating Federal Research Programs: Research and the Government Performance and Results Act* (1999; see <http://books.nap.edu/html/gpra>) and *Implementing the Government Performance and Results Act for Research* (2001; see <http://books.nap.edu/html/gpra2>) — and by other observations from the scientific community, such as the US House of Representatives Committee on Science 1998 report *Unlocking Our Future: Toward a New National Science Policy* (http://www.house.gov/science/science_policy_report.htm). The administration's efforts will start with applied research programmes, and apply the lessons learned to the evaluation of basic science. The administration supports explicit criteria for the quality, relevance and appropriate federal role of projects, rather than basing funding levels on vague anecdotes and previous-year funding levels.

The administration does not believe all basic research should be done by industry, nor are we developing criteria in isolation with merely 'green-eyeshade' views of research investments. We fully understand the difficulties of applying performance measures to basic research programmes, but this does not exclude good management and high performance. We have had extensive consultations with COSEPUP, the Office of Science and Technology Policy, members of Congress and their staff, and other leading researchers. These have provided useful insights that we are incorporating into our work.

The National Academy of Sciences has laid a foundation for us. We invite the scientific community to continue to work with us as we tackle the difficult task of implementing these sound ideas.

Mitchell E. Daniels, Jr

Director, Office of Management and Budget, 725 17th Street NW, Washington, DC 20009, USA

Echoes from the dreamtime

Analysing hunter-gatherer societies past, present and might-have-been.

Constructing Frames of Reference: An Analytical Method for Archaeological Theory Building Using Ethnographic and Environmental Data Sets

by Lewis R. Binford

University of California Press: 2001. 583 pp.
\$75, £52

Robert L. Bettinger

Ten thousand years ago, as the last ice age ended, the world was the exclusive domain of full-time hunter-gatherers — people who fed, clothed and sheltered themselves entirely with the fruits, seeds, meat, fibres and skins obtained by fishing, hunting and gathering wild plants. They are all gone now, but the handful who managed to hang on into the nineteenth and twentieth centuries figured prominently in the development of a modern anthropology whose goal of understanding all humankind demanded that special attention be given to these peoples, so profoundly out of step with civilization. As opportunities for first-hand study dwindled, however, anthropology traded holism for contemporary relevance and increasingly diverted its attention from hunter-gatherers to the national and transnational forces that were relentlessly putting them out of business.

But the subject matter would not go away. Archaeologists around the globe continued to excavate hunter-gatherer sites and, with the rest of anthropology enmeshed in modernity, proceeded to develop the theory needed to interpret them. Lewis Binford is the acknowledged leader in this field, and *Constructing Frames of Reference* is his latest and most extensive effort.

Long-time critic of interpretations in which archaeological remains are explained by matching them to a known ethnographic behaviour, Binford argues that the real question is not whether the archaeological and ethnographic cases are similar, but why they are similar. That is, what causes the behaviour they share? In this view, these behavioural patterns are the primary interest. Ethnography and archaeology merely provide instances of their particular expression in equally particular times and places. It follows that ethnographic hunter-gatherers are as much in need of explanation as archaeological ones, and in this book Binford mainly attends to the ethnographic part of the record, using comparative, cross-cultural analysis. Adding to a set compiled by G. P. Murdock, Binford has assembled key environmental and cultural variables for a sample of 339 hunter-gatherer groups scattered around the world.



Looking to the past: a people profoundly out of step with civilization.

As in his earlier work, Binford assumes that almost everything hunter-gatherers do is strongly conditioned (although perhaps not wholly determined) by the natural environment. He searches for relationships between nature and behaviour (for example, between latitude and technology), as well as for those between different aspects of nature (for example, latitude and natural productivity) and between different aspects of behaviour (such as family size and polygyny), in a dizzying array of maps, tables and bivariate scatter plots. As in all comparative research, the idea is to reveal forces that are at work, but impossible to detect in individual cases, by observing the interaction of variables across many cases. Because a major theme of the book is hunter-gatherer adaptive intensification — the trajectories of behavioural change that attend population growth — many of these comparisons investigate the relationship between subsistence, settlement and socio-political organization on the one hand and demography on the other. No amount of comparison will mean much, of course, unless the sample used is representative of the larger population one wishes to understand.

Binford is acutely aware that, if the larger population in question is all hunter-gatherers — the hunter-gatherer way of life in general (presuming there is such a thing) — his ethnographic hunter-gatherer sample falls well short of the mark. It is both geographi-

cally and behaviourally biased. It omits, for instance, much of Eurasia, where agriculturalists displaced hunter-gatherers thousands of years before anyone thought of doing anthropology, and it comprises only recent hunter-gatherers, groups at the end of a long trajectory of growth and adaptive intensification that began at the end of the Pleistocene.

It is possible, of course, to say some interesting things about recent hunter-gatherers from the ethnographic sample, but the geographical bias will always preclude broad theoretical generalization. Also, some of the most interesting things to be said about recent hunter-gatherers have less to do with the way they are than how they came to be that way — on which the ethnographic sample is largely mute. In consequence, Binford constructs two of his own hunter-gatherer samples, moving at this point beyond ethnography to the dimmer realm of hypothetical hunter-gatherers, who — like the revelations of the Australian Aboriginal dreamtime — do not exist in the present but serve to inform it.

The first sample is designed to represent what recent hunter-gatherer subsistence would have looked like without agriculture in places such as Europe, for instance, where hunter-gatherers were not ethnographically recorded. The second is designed to represent hunter-gatherer subsistence in the terminal Pleistocene, before 10,000 years ago and before population growth and intensification, when Binford assumes that

PETER JOHNSON/CORBIS

hunter-gatherers subsisted on terrestrial plants and animals almost to the exclusion of fish and aquatic mammals. Comparison of the two hypothetical samples thus provides clues about changes in the hunter-gatherer diet as population increased and traditional resources came under pressure. As the Pleistocene model excludes aquatic mammals and fish, their use naturally emerges as one signature of intensification, especially in Europe, where the postulated change is often from terrestrial animals to aquatic resources. In South America and Africa, on the other hand, the postulated change is mainly from terrestrial animals to terrestrial plants.

Binford draws several interesting insights from these comparisons, especially about the spread of agriculture into Europe, which he suspects was largely not the result of migration but rather a subsistence shift — one that occurred almost instantaneously in those places where hunter-gatherer populations had reached a critical point and needed to intensify, but lacked the wild resources to do so.

In the remaining chapters, Binford pursues hunter-gatherer intensification and the shift to agriculture in depth, repeatedly searching for patterns that reveal critical points, or thresholds, where a directional trend in population density, group size, or the like, is reversed. For him, these signal trajectories of intensification in which hunter-gatherer systems restructure radically from within, and are likely to produce archaeological signatures of discontinuity that Binford thinks are too often attributed to migration and ethnic replacement.

He is clearly not afraid to speculate and

take risks, and constantly pushes the edge of his interpretive envelope, never more so than near the end of the book. There he takes the climatic algorithms developed to project his hypothetical Late Pleistocene and ethnographic hunter-gatherer samples, applies them to detailed palaeoclimatic information for the Near East and North Africa, and uses the results to interpret the changes in behaviour, including the origin of agriculture, that took place in the terminal Pleistocene and the early Holocene that followed it.

This approach assumes, of course, that hunter-gatherers not only adapt to changes in climate (which seems reasonable), but that this adaptive readjustment more or less keeps pace with climate change. This is altogether more problematic, especially in light of ice-core records of wild and rapid climatic shifts, from full glacial to full interglacial conditions within the span of a few decades, during most of the Pleistocene — including the Late Pleistocene interval that interests Binford here. Abrupt termination of this chaotic pattern of climate change with the onset of the Holocene 10,000 years ago raises the further spectre of a potentially significant behavioural gulf in the range and scope of responses to climate change separating hunter-gatherers before and after this.

It is important to understand what this book is and is not. It is an important landmark work, but it is not for the general reader. Although it contains things of potential interest to geographers, ecologists and biologists, it is a technical volume written mainly for students of hunter-gatherers and will challenge the patience of even the most

quantitatively inclined of them. This is largely because Binford is less interested in using quantitative data to demonstrate statistically significant relationships than in using them to discover interesting and unsuspected patterns, which may or may not be convincing to readers without his predilections.

In contrast to many of Binford's earlier works, *Constructing Frames of Reference* contains very few simple take-home lessons. In compensation, he gives much greater attention than he has in the past to hunter-gatherer social and political organization and the ways in which they might be manipulated to facilitate intensification. The environmental and behavioural information Binford compiles for his sample of 339 hunter-gatherer groups will certainly be one of this book's most lasting contributions. Its theoretical and interpretive impact, although sure to be large, is harder to judge, simply because of the sheer scope of what is attempted, the wide subject matter covered and the data presented, all of which will take time to sort out. ■

Robert L. Bettinger is in the Department of Anthropology, University of California, Davis, California 95616, USA.

Biographical biology

Darwin & Co.: Eine Geschichte der Biologie in Portraits (in German)
edited by Ilse Jahn & Michael Schmitt
C. H. Beck: 2001. Volume 1: 552 pp. DM68.
Volume 2: 574 pp. DM68

Claus Wedekind

Who still believes that the theory of evolution was developed by Charles Darwin alone, uninfluenced by other ingenious thinkers? Progress is not made like this, only myths. But who can remember the names of those other biologists or their contributions to evolutionary theory, let alone any biographical details? This is now remedied in the more than 1,000 pages of the two volumes of *Darwin & Co.*, which contain 51 biographies of researchers who have shaped biology over the past three centuries. Given the title and the weight, one might fear that the result would be a tiresome account of errors and corrections, of misunderstandings and clarifications, around a single theory. But these fears are unjustified.

Darwin was surely a most influential biologist. Much has been written about him, his life, his family, the struggle with his contemporaries and, of course, his achievements. But in this book, he gets only the 21 or so pages allotted to each of the other personalities. The biographies are arranged chronologically, are not centred around evolutionary theory and can be read in any order.

Lost to the world

Only two whole specimens of Australia's short-tailed hopping-mouse (*Notomys amplus*) were found before the species became extinct in 1896. *A Gap in Nature: Discovering the World's Extinct Animals* by Tim Flannery and artist Peter Schouten (Atlantic Monthly Press, \$34.95) catalogues the 103 species of mammal, bird or reptile known to have become extinct between 1500 and 1999.





Strong ideas: the Comte de Buffon believed variation within a species was due to a *force pénétrante*.

Some of the authors are professional historians of science; others are biologists, often now retired, who happen to have an interest in the person. Despite this diverse authorship, the editors have managed to achieve a consistent structure for the portraits. Each contains a section on the subject's life story, embedded in the circumstances of the respective period and location, and often spiced up with subjective assessments of personality. The rest is a summary of the scientific achievements, largely well explained for non-specialists.

The collection starts with Carl Linnaeus (1707–78), who introduced a still-valid nomenclature for classifying organisms. As well as enormous productivity, Linnaeus possessed an ingenious imagination. As a physician, he suggested that infectious diseases such as smallpox, measles, dysentery, syphilis and plague might be caused by very small organisms. This was long before microscopes enabled us to see these pathogens. The idea must have sounded quite bold to his contemporaries, who used fragrances as protection against infection.

It was perhaps as bold as some other early biologists' ideas sound to us now. Georges-Louis Leclerc, Comte de Buffon (1707–88), another polymath of impressive productivity, explained variation within a species by what he called the *force pénétrante* and the following model. Organic particles that are transported by the blood system to the

different organs are either used for growth and maintenance, or, if they do not spontaneously get assembled into gut worms, get imprinted by these organs under the influence of the *force pénétrante*. They are then collected in seminal fluids. At coition, male and female seminal fluids are mixed and form offspring. These offspring resemble their parents because they are formed by imprinted organic particles. Later, the offspring's own organic particles will be modified as a result of different habits and environments. Hence — variation in time and space. It seemed as if you only had to explain the *force pénétrante* to complete the model.

Biologists of the twentieth century are, of course, 'just' specialists. But even if you are not that interested in agronomy, you would find the story of the botanist Nicolai Vavilov remarkable. He and some of his colleagues did not bow to lysenkoism and the threats of the stalinistic reign of terror in the Soviet Union of the 1930s and 1940s, and continued to discuss mendelian genetics and plant breeding. Because their opponents had political influence, Vavilov and his colleagues knew they risked arrest and death, and eventually paid this high price for their ideals.

Equally, even if you know little about ethology, you might be curious to learn how one of its founders, Konrad Lorenz, stood by his membership of the National Socialist Party of Nazi Germany, and how, as a

scientist, he arrived at his ideas about the preservation of pure races. Another example of free speech in science?

The biographical portraits end with the American geneticist Barbara McClintock, who died nine years ago. In passing, we learn that this excellent student and later Nobel prizewinner was not accepted for a PhD position at the Institute of Genetics at Cornell University in the 1920s because, at that time, this degree was not open to women.

Talent and diligence alone never guaranteed professional success. The stories in this book tell us that biologists have always needed influential supporters, and this is where scientists who have reached positions of authority can do some good. But who will remember those enthusiastic and talented students who came to grief against the pride or morbid ambition of some of these authorities? With this in mind, one might take a sceptical view of overly laudatory exaltations of the 'great man'. Indeed, some of the portraits have rather too many unnecessary superlatives. Nevertheless, what biologist would not be curious to read about the lives of their successful predecessors? I certainly enjoyed dipping here and there into these biographies, which are of such an easily digestible length. ■

Claus Wedekind is at the Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK.

More on Darwin

Charles Darwin: Interviews and Recollections

edited by Harold Orel
St Martin's Press, \$55

Darwin's Mentor: John Stevens Henslow, 1796–1861

by S. M. Walters & E. A. Stow
Cambridge University Press, £40

Benefactors of mankind

Cultures of Creativity: The Centennial Exhibition of the Nobel Prize

edited by Ulf Larsson
Science History Publications: 2001.
228 pp. \$40

The Nobel Prize: The First 100 Years

edited by Agneta Wallin Levinovitz & Nils Ringertz
Imperial College Press/World Scientific
Publishing: 2001. 236 pp. £18, \$26 (pbk)
Alison Abbott

Alfred Nobel was a tormented genius in his own right. A brilliant entrepreneur, by the age of 40 he was already making a fortune



Nobel (clockwise from left): Linus Pauling with his wife Ava Helen; Niels Bohr (left) holding court in Copenhagen; and Alfred Nobel's will.

from his inventions in the explosives business. But he had also witnessed the death of his brother in an explosion at the family's Stockholm factory — one of many deaths in Nobel factories around the world with which he had to cope — and had suffered a string of failed love affairs.

Nobel saw himself as both misanthrope and philanthropist, and referred to himself as a 'super-idealist'. As such, and to the dismay of his relatives, he left his fortune for the establishment of the Nobel prizes, to be awarded to those "who have conferred the greatest benefit to mankind". In celebrating the centenary of the first awards, the Nobel Foundation and the Nobel Museum have published commemorative books.

The Nobel Prize offers a brief history of the foundation and the research areas rewarded, and an explanation of how the prize evolved. 'Benefit to mankind' is, after all, a rather open term, and interpretation of Nobel's intentions was not straightforward. The book also explains how winners are selected — by the Royal Swedish Academy of Sciences in the case of physics and chemistry and by the Karolinska Institute in the case of physiology or medicine — and provides the occasional insiders' view on errors of inclusion (for example John Macleod, whose role in the discovery of insulin is now questioned) or omission (for example Oswald Avery, who identified DNA as the genetic material).

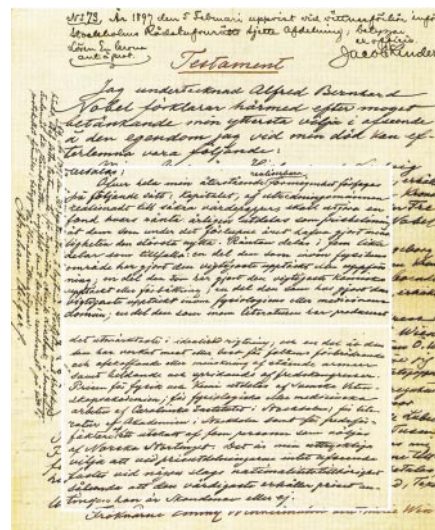
But the book disappoints. The quality of the essays is variable and the true insights to be found among the extended lists of prizewinners are dissatisfyingly few.

On a different tack, *Cultures of Creativity*, which accompanies an exhibition at the

Nobel Museum, ponders the essence of creativity in 44 sketches of selected laureates — but it also disappoints. According to Joseph Brodsky, winner of the Nobel prize for literature in 1987, "creativity is an unending exercise in uncertainty". But even in literature, the source of creativity is hard to pin down — in science even more so.

Perhaps Louis Pasteur came closest to the truth when he commented, 50 years before the first Nobel prizes were awarded, that "in the field of observation, chance favours the prepared mind". That is, to be creative means to know your subject inside out, the better to make connections between unexpected events, or to interpret unexpected results. Aside from confirming that hard work is a prerequisite for creative genius, *Cultures of Creativity* provides only superficial hints — and sometimes no hint at all — about the additional ingredients that make some hard-working scientists excel.

Still, the hints are amusing. Erwin Schrödinger (Physics, 1933), who apparently used pearls as earplugs to give him the quiet he needed for his research, did his best work when his private life was in turmoil. Linus Pauling (Chemistry, 1954, and Peace, 1963) called on his subconscious mind to chisel away at problems and provide him with inspiration long after the problems had been posed. August Krogh (Medicine, 1920) swore by "visual thinking", lying in bed. His ideas came out of the blue, but were worked through and validated inside his head,



without using a pen, which blocked the free flow of his thoughts. George de Hevesy (Chemistry, 1944) gained inspiration from the constant travel forced upon him in fascist Europe, which brought him into contact with many laboratories.

Cultures of Creativity also asks if there is such a thing as a creative place. The gold standard for creativity is acknowledged to be pre-war Copenhagen, where Niels Bohr (Physics, 1922) played eternal host to an intellectual elite — including many future Nobel prizewinners — who gained from endless discussions, and seeded throughout Europe the 'new physics' of the twentieth century. Postwar, this formula of encounter and discussion has been repeated successfully at several hothouses, including Cold Spring Harbor in New York and the Basel

Institute for Immunology. But if Max Perutz (Chemistry, 1962) is to be believed, discoveries cannot be planned in this way — “they show up by themselves in unexpected places”. Nonetheless, Perutz was wise enough to build a gourmet canteen in his research institute in Cambridge, a city that lays claim to 70 laureates, to encourage scientists to meet up and talk. ■

Alison Abbott is senior European correspondent of Nature.

Yes, I remember it well ...

The Seven Sins of Memory: How the Mind Forgets and Remembers
by Daniel L. Schacter

Houghton Mifflin: 2001. 336 pp. \$25, £17

Larry R. Squire

The comedian Alan King tells the story of being invited by his wife to join her and some friends at a newly opened Italian restaurant. “A new restaurant?” he replied incredulously. “It took me 20 years to find five restaurants I can trust.” In any bookstore, on the psychology and self-help shelves, one can find dozens of books about the mind and the brain, including books on such popular topics as sleep, dreams, mood, thinking and creativity. Some are based on testimony or personal



Lining up the usual suspects: but our memory has a tendency to select the person most like the culprit.

experience; others purport to be based on science. Is it possible to navigate through this material and find information that one can trust? There is Thomas Gilovich's *How We Know What Isn't So* (reprinted by Free Press, 1993), an entertaining and example-filled treatment of the kinds of cognitive errors that commonly attend evidence gathering and data interpretation. Similarly, in *Inevitable*

Illusions (Wiley, 1994), Massimo Piattelli-Palmarini describes well-studied illusions of probability and judgement, ways in which human reason is prone to error in everyday life.

In the same tradition, in *The Seven Sins of Memory* Daniel Schacter has written about the frailties of memory, describing the multitude of ways in which it is prone to selectivity, distortion and all-out failure. The fallibility of memory is not itself news. Jokes about memory are part of the culture, and people beyond the age of 40 often seem to regard their memories as flawed, unreliable and perhaps even worthy of study. It emerges, though, from laboratory studies of memory and from Schacter's analysis of seven kinds of memory failure (the “seven sins”), that errors of memory are inherent in its nature — quantifiable, reliable and universal aspects of cognition. The book makes the additional argument, although this point is harder to settle, that memory's imperfections are by-products of adaptive design.

Memory is not a single thing, and many of its familiar shortcomings are quite unrelated to one another. For example, simple forgetfulness, such as forgetting what one heard (the first sin), depends largely on how deeply information is processed at the time of learning, how much it is rehearsed and whether it can be related to what one already knows. The important brain regions are the medial temporal and inferior frontal lobes. After learning is completed, the course of forgetting is related to the gradual loss of stability within relevant neural circuits. In contrast, the common experience that the name of a familiar person may be momentarily inaccessible (blocking, the third sin) has a separate neural substrate. Persistence (the



Legacy of the Incas

The Intihuatana, or Hitching Post of the Sun, is the best known of the many sacred carved rocks and shrines of Machu Picchu, the ancient Inca ruins that lay hidden for 400 years until they

were discovered by the Yale University archaeologist Hiram Bingham in 1911. Taken from *Realm of the Incas* by Max Milligan (HarperCollins, £29.99).

KOBAL COLLECTION/CHEN, LINDA R./POLYGRAM/SPELLING

seventh sin) is different altogether. Persistence refers to the retention of highly emotional memories that one would rather forget. This independence of mechanism means that a treatment, such as a drug, developed to help 'memory' is unlikely to work on all of memory's imperfections.

Schacter's framework of research findings and interpretations, organized around his seven categories, succeeds in describing a discipline of work, theoretical and applied, that is both fascinating and of practical importance. The tip-of-the-tongue phenomenon (the feeling that a momentarily unavailable word or name is on the tip of the tongue) is experienced so similarly across cultures that 45 of 51 languages surveyed use expressions containing the word 'tongue' to describe this psychological state (including Estonian, Cheyenne and Afrikaans). What is commonly termed absent-mindedness (the second sin) is due to a disconnection between attention and memory, for example, a failure to attend at the moment when information is encoded. Absent-mindedness also occurs when retrieval cues are not available or are not noticed when an intended action is meant to be carried out.

Another imperfection involves the tendency to misremember the past so as to make it more consistent with current knowledge and beliefs (bias, the sixth sin). College students who were dating were asked about their partners on different occasions two months apart. Those whose evaluations become more negative during the two months recalled their original impressions as more negative than they actually were. Conversely, those whose evaluations became more positive recalled their original impressions as more positive than they actually were.

During the past decade, the fragility of memory became a prominent issue in the courtroom and in the news. Here, the guilty features of memory are misattribution (the fourth sin) and suggestibility (the fifth sin). Misattribution refers to errors in which a thought or a dream or a general feeling of familiarity is reported as a real memory. Suggestibility refers to distortions of memory through leading questions or other forms of influence. It is easy to demonstrate in the laboratory how readily individuals will succumb to memory errors of these kinds. After seeing a word list (bed, rest, awake, tired, dream, wake, snooze, blanket, doze, slumber, snore, nap, peace, yawn, drowsy), individuals frequently report with confidence that a highly associated word (sleep) was on the list. Persons with a capacity for vivid imaging, who score high on suggestibility scales, show a greater tendency for this effect.

There are many examples where misattribution and suggestibility appear to be at work. These include the corruptibility of young children's memories in the face of

leading questions, false confessions, early memories of abuse that were 'recovered' during psychotherapy (and later disavowed), and eyewitness misidentification. It is heartening to see that newer studies reviewed by Schacter are having practical consequences. For example, line-up identification procedures have been improved in some places by the simple expedient of presenting suspects one at a time and asking for a yes-no judgement. In the traditional method, the observer makes a choice after seeing all the suspects, but is vulnerable to error when the culprit is not in the line-up, because of a tendency to select the person who looks most like the culprit.

This book brings together an enormous amount of material in a form that is entertaining, informative and credible. It can both be enjoyed by the general reader and used as a resource by the specialist, as the text is accompanied by 19 pages of notes and citations supporting its empirical statements. Lawyers, judges, health-care professionals and others whose work sometimes relies on intensely practical questions about memory should also find it a welcome source of information. ■

Larry R. Squire is at the Veterans Affairs Healthcare System and in the Departments of Psychiatry, Neurosciences and Psychology, University of California, San Diego, California 92161, USA.

A physical view of the cytoskeleton

Mechanics of Motor Proteins and the Cytoskeleton

by Jonathon Howard
Sinauer/Palgrave: 2001. 384 pp.
\$59.95/£44.99

Edwin Taylor

The cytoskeleton consists of three kinds of fibrous protein filament — actin filaments, intermediate filaments and microtubules — plus the proteins that interact with them. The molecular motors myosin, kinesin and dynein run along the filament tracks. At least 100 other proteins interact with the cytoskeleton and new ones continue to be discovered.

The cytoskeleton is an area of intense research and we are in danger of drowning in a sea of facts. What should we try to teach our students about it? Popular textbooks devote one or two chapters to the cytoskeleton and motility and they can be supplemented by a book such as Dennis Bray's *Cell Movements* (Garland, 2001). But a textbook is needed which starts from first principles and leads to an understanding of the dynamics of the system.

And here is that book. In the first part, Joe Howard presents the physical principles that govern the behaviour of the cytoskeleton. It is a mechanical system composed of thin rods that are under stress. Movements occur in a medium in which inertia is negligible compared to viscous damping. The basic problem is the motion of a small elastic body in a viscous medium subject to an applied force and thermal motion. This problem is presented in a clear, conversational style and the level is that of a first-year college physics course in which some calculus is needed.

It may seem surprising that important properties of fibrous proteins and motors can be deduced from their resistance to stretching and bending, but proteins are large enough to be treated as bits of continuous matter. Howard points out that the elastic properties of these proteins have very little to do with the amino-acid sequence or even with the three-dimensional structure, because the proteins are held together by the same kinds of weak bonds.

Ask a student or colleague how long it takes a protein of relative molecular mass 100,000 to diffuse 1 millimetre, or what is the rotational time of a myosin crossbridge (or any other protein of similar shape). The chances are that the answers will be wrong by a few orders of magnitude. By including many simple calculations, some left for the reader to do, Howard succeeds in building up a feeling for the dynamics of the cytoskeleton.

The results derived in the first part of the book are then used to treat the mechanical properties of the cytoskeleton. The molecular motors myosin and kinesin are discussed in the third and final part. Howard is a major contributor to studies of kinesin, and provides a clear description of the techniques of mechanical measurements on single molecules.

This book would be an excellent text for a semester course on the cytoskeleton and motility. Most of the topics are accessible to students who have had a course in physics. Some topics in diffusion and thermal motion use more advanced mathematics, but the solutions to equations have been wisely placed in an appendix where they don't get in the way of the reader who is interested in results, not mathematical derivations. It will also be a valuable reference book for researchers in the motility field.

Students on general cell-biology courses often have an aversion to equations, but a selection of some of the results discussed in this book would be a valuable supplement to such a course. ■

Edwin Taylor is in the Department of Molecular Genetics and Cell Biology, University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA.

Science's fiction

Scientists deal with the facts. But they wouldn't get anywhere without dreaming up stories first.

Jean-Marc Lévy-Leblond

Fact versus fiction — is there a more entrenched opposition? It underlies our very conception of science, which investigates solid facts of the world, in contrast to art and poetry, which produce fictitious creations of the mind. Indeed, the very word 'fiction' comes from the Latin verb *fingere* , meaning to fake, to feign, to pretend. This is exactly how Newton used it when he claimed "*Hypotheses non fingo*" to express his refusal to contrive an ad hoc mechanism for explaining his law of gravity. "I stick to the facts" is perhaps the most faithful translation, an assertion repeated by many physicists since.

Rejecting fiction is meant to insure the positivity of scientific knowledge against the risk of uncontrolled imagination. Yet the drafting of hypotheses — that is, precisely, *fictions* in Latin — is one of the first endeavours of scientific activity. How, then, can we discriminate between hypotheses that we should dismiss as irrational fantasies, and those that we posit right at the outset of our investigations? Could it be that science paradoxically offers the best proof that fiction can lead to facts?

Let us go back to our etymological inquiry, which we interrupted too soon. In fact, the verb *fingere* , in archaic Latin, is very concrete and factual — its initial meaning is to model (in clay), to sculpt, to mould, to represent. The word thus forms the root of 'fictions' and 'figments', but it is just as relevant to 'effigies' and 'figures'.

It is the very history of language, then, that advises us not to oppose fictions, as inventions and creations, to figures, as representations and models. The former are to the latter as imagination to imaging, a deployment without discontinuity. It now becomes possible to think that science and fiction are not incompatible — in spite of the purposely oxymoronic use of the two words in the naming of the literary genre known as science fiction. I would even advocate the idea that theoretical physics, that paragon of exact sciences, is, first and foremost, fiction. The physicist, as does the novelist, invents worlds and tells stories. Any historical episode of some import illustrates this thesis.

Euclidean geometry, which is a physics of space, already deals essentially with the basic geometrical figures, which are but fictions — points with no extension, lines with

We will not shun exploration of hypothetical universes in case our own world should resemble them.

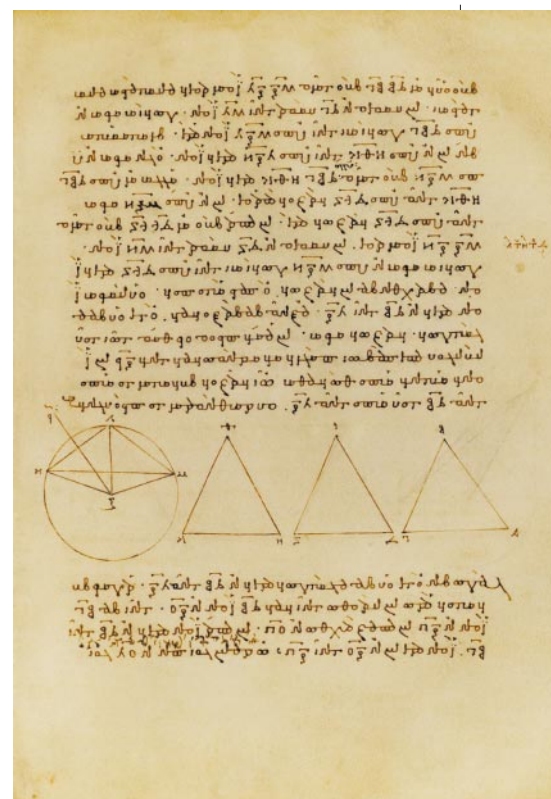
no end, planes with no width. These first mathematical entities have no factual existence, and are (worthwhile) figments of the imagination.

Galilean physics gave us the law of free-fall, which is mother to all legislation of the physical world. But its claim of constant acceleration cannot be true except in a perfect vacuum. This statement thus starts as a children's game: "Pretend there is no air ...", to which one must immediately add "... and pretend that the Earth is flat and does not move", and so on. Hypotheses, indeed, that we feign and try fitting to the facts.

Contemporary theoretical physics is no different. Nuclear forces are studied as if gravity did not exist. Special relativity describes the structure of space-time as if it were empty. And, following Einstein, the recourse to a *Gedankenexperiment* (fictitious experiment) is one of the favourite methods of modern theoreticians.

These are instances in which fiction operates by restriction, so to speak, letting us study simplified figures and reduced models of the world. But there are other cases in which it operates by extension, leading us to investigate radically different universes, reaching well outside our representations. To stay within theoretical physics, although our space-time has 3 + 1 dimensions, we will not shun exploration of hypothetical universes with 10 or 26 dimensions, in case our own world should resemble some corner of these super-universes. Although a powerful and sturdy theory forbids any object from exceeding the limiting velocity of light, we will dare to conceive of 'tachyons', supraluminous particles, and then desperately try to clear up apparent contradictions in their behaviour. The scientist is an unrepentant dreamer — far from sticking to factual observations, he must imagine fictitious situations, which may, from time to time, prove to be veracious.

Such a thesis might be accepted without too much reluctance as far as it concerns the



Art imitating life: the fictitious constructs of euclidean geometry help us to describe reality.

theoretical activity of science. But, it will be objected, this cannot hold for science's other facet, experimental practice, through which science grapples with external reality so brutally that friction, rather than fiction, becomes the password. Indeed, far be it from me to deny or underplay this confrontation, which permits the specificity of scientific knowledge.

Yet no proper experiment is a direct and naive affair with nature. What are sophisticated experimental apparatuses, if not fictional devices that enable the outside world to reveal itself in comprehensible terms? Why would we need to experiment, that is, to imagine and produce artificial phenomena, if mere observation of natural ones sufficed? Forcing bodies to fall along a slanted trajectory instead of letting them follow the spontaneous vertical direction (Galileo); coercing electrons to circulate along manmade metallic wires (Faraday, Ampère); and creating *de novo* elements endowed with artificial radioactivity (Irène and Frederic Joliot-Curie) — are these not tantamount to requiring nature to tell us unheard stories and to unfold new narratives?

According to Jean Cocteau, poetry is "a lie which tells the truth". The same is true of science. At least, this seems an interesting hypothesis to feign.

Jean-Marc Lévy-Leblond is in the Departments of Physics and Philosophy, Université de Nice, Parc Valrose, F-06108 Nice cedex, France.

Unweaving the whirls

Joël Sommeria

Turbulence denotes complex fluid motion, as opposed to smooth laminar flows. It evokes agitation, disorder and chaos. The definition of chaos restricts it to dynamic systems of which only the time behaviour is complex — deterministic but sensitive to the initial condition. Turbulence involves spatial as well as temporal complexity, with 'fractal' multi-scale velocity fields.

Although physics is successful in grasping many remote and extreme phenomena, it is often of little help with common problems encountered in life. In *De rerum natura*, Lucretius describes the random motion of dust seen in a ray of sunlight, an early hint at the motion of atoms in a vacuum. Yet nowadays, we know more about atoms than about the turbulent air motion revealed by dust.

The relevant laws of mechanics have been known since Newton's time; the resulting partial-differential equations for a fluid were derived by Leonhard Euler, with a viscous term introduced by Claude-Louis Navier in 1823. Turbulence is supposedly included in these equations as complex, unsteady flow solutions. Modern computers have begun to grasp this complexity, but the available spatial resolution is still too low for a full model of the air flow in a room. We need to resolve the scale on which viscous smoothing occurs

— just a few millimetres — and we also need to determine the initial condition. But measuring such three-dimensional velocity fields is beyond the ability of current technology.

A direct numerical simulation would provide gigabytes of raw data, but this would just be the starting point. Statistical properties, such as mean transport of air mass, of temperature and of chemicals, must be considered. Wall friction and noise emission (pressure fluctuations) are relevant in aerodynamics. The gradients of wind and concentration of transported chemicals have intriguing, intermittent fluctuations, which can control chemical reactions. The emergence of coherent flow structures is also remarkable, but is often difficult to characterize unambiguously.

'Predictability' expresses the sensitivity of predictions to a small initial perturbation, or error — thereby imposing, for example, limitations on meteorological forecasts. Exponential error growth (chaos) is observed in simplified models that retain only the largest scales of motion. (This has been popularized as the 'butterfly' effect, from the legend about a butterfly wingbeat eventually modifying the weather.) This is unlikely, however, as such unpredictability is overestimated in models that are simplified to a few dynamic variables describing only large scales. In reality, the global effects of many small-scale perturbations tend to be predictable. Likewise, colliding molecules have highly unpredictable motion, but collectively give rise to the very predictable properties of gases.

Sensitivity to physical perturbations (for instance, vibration at a wall) is also a key issue, about which laboratory experiments provide valuable information. Turbulence is therefore a genuine scientific field, involving experiments, theory and numerical simulations. Although it relies on classical mechanics equations, the solutions of which are approached by computation, conceptual understanding is needed. This is essential not only in engineering design — for instance in improving aerodynamics by active control of boundary layers — but also for modelling the evolution of our natural environment.

In the early Universe, turbulence controlled the development of density fluctuations that eventually led to the formation of galaxies. Similarly, it governs the development of fluctuations in molecular clouds at the origins of stars, as well as the accretion of protoplanetary nebulae and the agglomeration of dust that leads to planetary formation. Climate changes are governed by turbulent fluctuations in the Sun, the atmosphere and the oceans, and the melting of polar ice caps is controlled by turbulent heat transfer.

The concept of eddy viscosity (and diffu-

Turbulence

Turbulence is a genuine scientific field, involving experiments, theory and numerical simulations.

sivity) is widely used to model the effects of the small, unresolved scales of turbulence in 'large eddy simulations'. Navier derived his expression of viscosity as an effect of unspecified random transport; such arguments were later formalized by James Clerk Maxwell and Ludwig Boltzmann in the kinetic theory of gases. The statistics of a continuum fluid motion pose a quite different problem: unlike molecules, fluid elements interact over long ranges (by pressure force) and deform irreversibly.

In two dimensions, the vorticity (curl of velocity) of each fluid element is conserved. The system is therefore like a set of elementary vortices (spinning-tops) with a long-range interaction. Statistical equilibrium states involve clusters of aligned elementary vortices. This explains the persistence of large, coherent vortices, such as Jupiter's Great Red Spot, in turbulent atmospheres and oceans.

But the most common form of turbulence is three-dimensional and far from equilibrium, with an irreversible cascade of energy towards small scales. Andrei Kolmogorov described this cascade in 1941, arguing that the dynamics are scale-independent. However, this is only approximately true. Deviations have been obtained, in the form of intermittency exponents, for transport of a pollutant by a random gaussian velocity (simplified synthetic turbulence). Extension of these descriptions to turbulent flow itself would be a great achievement, but not the whole story.

Real turbulent flows are diverse, depending on boundary effects, rotation and external forces (such as buoyancy). Randomness is not externally imposed, but rather spontaneously arises. There will be no single breakthrough — we face deep questions about information processing and probabilities, and about the very meaning of scientific understanding and prediction. Turbulence allows us to venture towards this new frontier with the safeguard of rigorous classical mechanics. ■

Joël Sommeria is at the Centre National de la Recherche Scientifique laboratory LEGI/Coriolis, 21 Avenue des Martyrs, 38000 Grenoble, France.

FURTHER READING:

Frisch, U. *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge Univ. Press, Cambridge, 1995).
Lesieur, M. *Turbulence in Fluids* (Kluwer Academic, Dordrecht, The Netherlands, 1997).
Lesieur, M., Yaglom, A. & David, F. *New Trends in Turbulence* (Springer, Heidelberg, 2001).
Shraiman, B. I. & Siggia, E. D. *Nature* **405**, 639–646 (2000).



Up in smoke: the erratic motion of the particles reveals the turbulence of the surrounding air.

Semiconductors are cool

Cronin B. Vining

In the 1950s there were hopes that semiconductor thermocouples would replace mechanical refrigerators, just as semiconductor transistors supplanted vacuum tubes. New materials may bring that goal a bit closer.

Nothing so induces sleep as a lecture on thermodynamics. So we'll skip that part. But on page 597 of this issue¹ Venkatasubramanian *et al.* describe thin-film thermocouple devices made from new materials that could make all refrigerators and power generators in the world obsolete. Sounds too good to be true? Well, perhaps better picnic coolers at least.

Thermocouples based on metal wires are cheap, reliable and widely used for measuring temperature. A thermocouple is a simple electric circuit, formed by two dissimilar conductors joined at one end, that generates a voltage when the joint and the free ends are at different temperatures. But thermocouples can do more than just generate a voltage. They can operate as heat engines to convert heat into electrical energy (the Seebeck effect), or convert electrical energy into cooling (the Peltier effect; Fig. 1). Thermoelectric devices based on these effects could form a new generation of non-mechanical generators and refrigerators, if only they were more efficient.

Following the development of semiconductors in the 1950s, it was found that replacing the metal wires with semiconductors improved the efficiency of thermocouples by more than an order of magnitude. A big improvement, but normal thermoelectric technology would still cost too much and consume too much electricity to replace that compressor in your kitchen fridge. But semiconductor thermoelectrics are rugged, durable, solid-state energy-conversion devices, and have long been the technology of choice to provide power for deep-space missions, including the Voyager I and II probes to the outer planets and, more recently, the Cassini mission to Saturn. They are also well suited to certain remote, extreme environments (such as some oil pipelines) on Earth, and to small-scale cooling for defence and aerospace applications (such as cooling infrared detectors).

In the past decade or so, cost reductions have led to the introduction of thermoelectric (Peltier) coolers into consumer products, such as ice-less picnic baskets (putting cigarette lighters in cars to good use) and 'climate-controlled' (cooled or heated) car seats. Wristwatches powered exclusively by the heat from your wrist are also available, although they are expensive. But going beyond such niche markets requires much

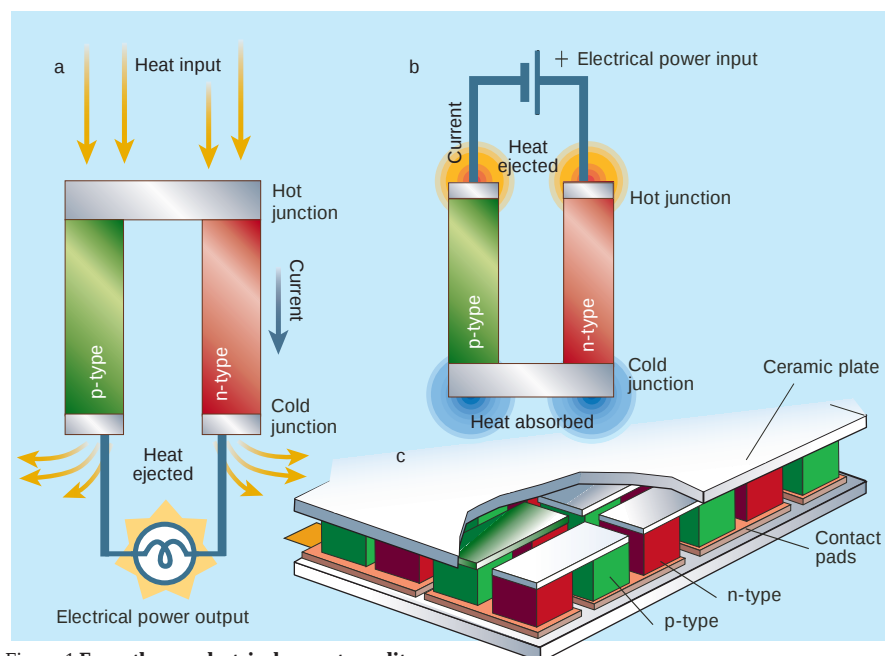


Figure 1 From thermoelectric dreams to reality.

Thermocouples are simple electric circuits used to measure temperatures owing to the voltage generated by temperature differences at the junctions formed by two dissimilar wires or semiconductors. They can also be used to generate: a, electrical power, or b, cooling. The efficiency of energy conversion is determined more by the properties of the thermoelectric materials (here the n-type and p-type semiconductors) than by the geometry. c, State-of-the-art thermoelectric devices can contain up to several thousand individual thermocouples. The electrical and thermal characteristics can be tailored to specific applications by adjusting the number of thermocouples in series and by varying geometric factors. The new thermoelectric materials developed by Venkatasubramanian *et al.*¹ are a major step towards more widespread use of thermoelectric technology beyond existing niche applications. (Graphics courtesy of S. Williams, www.thermoelectrics.com.)

better performance. The problem isn't in the engineering; it is with the science — in particular, the science of thermoelectric materials.

The problem is that the active materials themselves, the n-type and p-type semiconductor 'legs' indicated in Fig. 1, limit the efficiency. Electrons carry currents flowing in an n-type semiconductor, whereas the currents in a p-type semiconductor are carried by positively charged 'holes'. Both types have an electrical resistance, which must be overcome for the thermocouple to work. And each semiconductor leg conducts heat directly through the device, which limits the temperature difference a device can attain (for cooling) or maintain (for power generation). These losses detract from the ability of each leg to produce either a useful voltage or cooling, depending on the application.

How effective a material will be in a properly engineered thermocouple is measured by a combination of material properties known as the dimensionless thermoelectric figure of merit, usually written as ZT . Think of ZT as shorthand for 'efficiency'. ZT has the advantage that it can be measured on a single leg without building a full device. It can be zero, which means there is no energy conversion. As ZT increases towards infinity a thermoelectric device asymptotically approaches the usual Carnot efficiency limit, which applies to all heat engines (even solid-state ones).

Every material has a figure of merit, usually very small. The term 'thermoelectric material' probably should be reserved for materials with $ZT > 0.5$, which is rare enough to be interesting. For about 40 years, the best-known materials had ZT values

between about 0.75 and 1.0. That's why the work by Venkatasubramanian and co-workers¹ is so interesting: they report a ZT of 2.4 in thin films of $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ semiconductors. These materials appear to achieve such high ZT s thanks to their unusual structure — a superlattice formed by alternating layers of semiconductors. The previous record for ZT at room temperature was held by a bulk semiconductor alloy based on Bi_2Te_3 and Sb_2Te_3 . The superlattice structure appears to enhance the transport of current-carrying electrons (and holes) while inhibiting transport of heat-carrying phonons (quantized vibrations of the crystal lattice). Both effects boost ZT .

When the modern era of thermoelectric science and technology began to emerge in the late 1950s, it seemed possible that thermoelectrics might approach the efficiency of mechanical refrigerators and power generators. By the 1970s, given the lack of progress, few thought it likely. There was even speculation that a ZT of 1 represented some sort of thermoelectric barrier. Certainly it was an empirical limit that nearly halted research and development. But in the early 1990s Rudolph Buser, then associated with the United States Army Night and Electro-Optics Directorate, called on scientists to re-examine thermoelectrics. A basic science programme to increase ZT was soon underway, with support principally from the US Navy's Office of Naval Research and DARPA (Defense Advanced Research Projects Agency). By the late 1990s there was some progress, but even then you had to be an optimist to believe the barrier had been broken².

With the results of Venkatasubramanian *et al.*¹, even sceptics and dispassionate observers can safely be encouraged. The material properties, as measured by the figure of merit ZT , are 2.5 times better than the current state of the art, have been verified by more than one method, and are useful at room temperature. It has been a long time in coming but any conjecture about a thermoelectric barrier of $ZT = 1$ seems to have been safely put to rest.

Is it time to replace your old-fashioned fridge? Not just yet. As promising as these new results are, the efficiency (estimated from ZT) remains significantly less than that of conventional refrigerators. And there is no telling when, or if, costs and various engineering issues can be resolved.

On the other hand, this result may be good enough to greatly expand the range of practical applications. After all, modern manufacturers are good at reducing costs and there is no reason to believe this is the last word in efficiency. And most physicists can remember when the upper limit for superconducting transition temperatures was rather firm at about 23 K (-250°C), whereas the record now stands at 164 K (-109°C) — still pretty cold, but few would now bet against it going higher. Experimentalists just love to prove theorists wrong. ■

Cronin B. Vining is at ZT Services, 2203 Johns Circle, Auburn, Alabama 36830-7113, USA.
e-mail: nature@zts.com

1. Venkatasubramanian, R., Siivola, E., Colpitts, T. & O'Quinn, B. *Nature* **413**, 597–602 (2001).
2. Dubois, L. H. *18th International Conference on Thermoelectrics* 1–4 (IEEE, Piscataway, New Jersey, 1999); <http://www.zts.com/darpa/dubois99>

CO_2 from the atmosphere) and respiration (which releases CO_2 to the atmosphere) will increase with increasing temperature in a predictable way that will remain constant over time. Luo and colleagues¹ tested whether this assumption applies to soil respiration — the combined respiration of roots and micro- and macro-organisms in the soil. They used infrared heaters to warm plots of tall grass prairie by about 2°C over a period of one year, and compared these plots with unheated control plots. Although soil respiration was expected to increase by 15–20%, there was no significant change. The authors attribute this to respiratory acclimatization to the warmer temperatures: as temperatures rise, they suggest, the sensitivity of respiration to increased temperature decreases, thereby weakening terrestrial feedback to global warming.

A similar conclusion was reached recently by Xu and Qi³, who studied a forest in the Sierra Nevada. In their study, which took advantage of spatial and seasonal variations in soil respiration, the sensitivity of soil respiration to increased temperature was lowest in the summer, when temperatures were highest. But in this case it was difficult to distinguish between the effects of higher temperature and lower soil moisture.

These authors' emphasis on soil respiration is entirely appropriate because it constitutes the second largest pathway in the global carbon cycle, second only to gross primary productivity. Global estimates^{4,5} of soil respiration are in the range $68\text{--}100\text{ Pg C yr}^{-1}$ (Pg being petagrams, 10^{15} g); to put this in perspective, the annual input of CO_2 to the atmosphere through human activities is about 7 Pg C yr^{-1} . Evidently, then, even a small increase in soil respiration could accelerate climate change in the twenty-first century; conversely, a small decrease could compensate for anthropogenic emissions, and so slow the expected rate of change.

It has long been known that there is a strong link between soil temperature and soil respiration — respiration increases with rising temperatures, and vice versa^{5–7}. A standard way of defining this relationship is to calculate the Q_{10} value, which is the increase in respiration for each 10°C rise in temperature. Reported Q_{10} values⁵ for soil respiration are typically in the range 1.3–3.3, with a mean of about 2.4. But although the concept is a useful one, care is needed in using Q_{10} relationships to infer long-term trends: they are relatively simplistic, and are derived largely from laboratory experiments or short-term field studies of five years or less.

Unlike photosynthesis, soil respiration is not a single process. Rather, it is the summed activity of a complex and changing assemblage of below-ground organisms, including roots, microflora and micro- and macro-fauna. These organisms respond not only to

Global change

Matter of time on the prairie

Lindsey Rustad

In some ecosystems at least, extrapolating from the short-term effects of global warming will give a misleading impression of the reaction over longer periods of time.

The Earth is warming. Given that CO_2 seems to be the main determinant of global temperature, predictions of climate conditions in the future depend in part on gauging the response of the carbon cycle to warming. This is a question that Luo *et al.* address on page 622 of this issue¹. In their experiments on a terrestrial ecosystem in Oklahoma, they find that, in this case at least, acclimatization to increased temperature means that feedback of CO_2 into the atmosphere in the long term would be less than expected.

Over the past century, the Earth's mean surface temperature has increased by 0.6°C . During the next 100 years, it is predicted to increase by a further 1.4 to 5.8°C , and by even

more at higher latitudes². This predicted rate of change is unprecedented in at least the past 10,000 years, and is largely attributed to increases in the greenhouse gases, most notably CO_2 , resulting from the burning of fossil fuels and changes in land use. Confidence in these predictions is increasing, but considerable uncertainties remain. We can't even be sure whether terrestrial ecosystems will take up atmospheric CO_2 (and so moderate further increases in temperature), or be a source of it (and so drive temperature even higher).

In predictions of these climate–ecosystem interactions it is often assumed that, if moisture and nutrients are not limiting factors, rates of both photosynthesis (which removes

changes in temperature, but also to changes in moisture, nutrient availability and substrate quality, all of which are also directly or indirectly affected by temperature. For instance, the predicted relationships between soil temperature and soil respiration do not hold when moisture becomes limiting^{8–10}, or when there is a shift in the composition of the microbial community¹¹, or when there is a change in substrate quality or quantity^{12,13}.

The Harvard Forest soil warming experiment in Petersham, Massachusetts, provides a useful example of this latter case. In this study¹², the researchers found that soil respiration increased in response to experimental warming. Soil respiration was 40% greater in warmed plots than in control plots during the first year of the experiment, the rise probably being fuelled by the microbial oxidation of labile (easily decomposed) carbon compounds. The magnitude of the response of soil respiration to experimental warming declined markedly in the second year, however, presumably because the labile carbon supply was depleted. If a Q_{10} value from the first year of data had been used to extrapolate results and predict longer-term respiratory responses to warming, it would have resulted in a large overestimate of the amount of carbon released from the forest soil, and so of the potential feedback to climatic warming.

Luo and colleagues¹ likewise find a decline in the temperature sensitivity of soil

respiration with warming: the Q_{10} was 2.70 in the unheated plots compared with 2.43 in the heated plots. As with the Harvard Forest example, the results show that caution should be used in extrapolating results from short-term experiments to predict longer-term responses to environmental perturbations such as warming. The question of how ecosystems might or might not acclimatize to a warmer world bears serious consideration. But as with much research on this topic, longer time series of data will be needed to provide plausible answers. ■

Lindsey Rustad is in the US Department of Agriculture Forest Service, Northeastern Research Station, 271 Mast Road, Durham, New Hampshire 03824, USA.

e-mail: rustad@maine.edu

1. Luo, Y., Wan, S., Hui, D. & Wallace, L. L. *Nature* **413**, 622–625 (2001).
2. Intergovernmental Panel on Climate Change *Climate Change 2001: The Scientific Basis* (Cambridge Univ. Press, 2001); <http://www.ipcc.ch>
3. Xu, M. & Qi, Y. *Glob. Biogeochem. Cycles* **15**, 687–696 (2001).
4. Musselman, R. C. & Fox, D. G. *J. Air Waste Mgmt Assoc.* **41**, 798–807 (1991).
5. Raich, J. W. & Schlesinger, W. H. *Tellus* **44**, 81–89 (1992).
6. Rustad, L. E. *et al. Oecologia* **126**, 543–562 (2001).
7. Kirschbaum, M. *Soil Biol. Biochem.* **27**, 753–760 (1995).
8. Saleski, S. R., Harte, J. & Torn, M. S. *Glob. Change Biol.* **5**, 125–141 (1999).
9. McHale, P. J., Mitchell, M. J. & Bowles, F. P. *Can. J. Forest Res.* **28**, 1365–1372 (1998).
10. Rustad, L. E. & Fernandez, I. J. *Glob. Change Biol.* **4**, 597–607 (1998).
11. Zogg, G. P. *et al. Soil Sci. Soc. Am. J.* **61**, 475–481 (1997).
12. Peterjohn, W. T. *et al. Ecol. Appl.* **4**, 617–625 (1994).
13. Holland, E. A. *et al. Glob. Change Biol.* **1**, 115–123 (1995).

Ion channels

Swimming with sperm

David L. Garbers

Mice become infertile if they lack the gene encoding a newly discovered sperm-specific ion channel. Sperm are produced in normal quantities, but have trouble moving.

Spermatozoa rely on calcium ions to function. And, like the cells from which they are produced, sperm seem to express several channels that allow calcium ions to enter¹. But, so far, none of these ion channels has been linked to the regulation of sperm motility. That changes with an impressive paper by Ren and colleagues², published on page 603 of this issue. The authors have discovered a channel, which probably allows calcium ions to pass through, that is expressed only in sperm and is needed for them to move normally.

When sperm are first produced in the testes they are immobile (Fig. 1, overleaf). It is only after they are moved to their storage site, the epididymis, that they acquire the ability to swim forwards (progressive motility) — a behaviour that is required for successful fertilization. Sperm do not actually move about in the epididymis, but

actively swim forwards after ejaculation or dilution into various media. As they enter the isthmus of the female reproductive tract, sperm slow down once more³. They resume their migration when ovulation occurs, eventually reaching the ampulla region of the oviduct, where fertilization takes place.

As well as acquiring the ability to move progressively, sperm must undergo a further maturation process, termed capacitation, before they can fertilize an egg. This occurs while they are in the female reproductive tract, and results in two changes in sperm behaviour. First, they become able to undergo an acrosome reaction in response to the egg's extracellular matrix (zona pellucida), which involves the release of matrix-digesting enzymes. Second, sperm motility is hyperactivated.

Cyclic nucleotides, Ca^{2+} ions and intra-

cellular pH have been all proposed to regulate progressive motility and the events associated with capacitation, including the change to 'whiplash' hyperactivated motility³. Spermatozoa express voltage-gated Ca^{2+} channels, cyclic-nucleotide-gated channels and transient receptor potential channels (a different type of putative Ca^{2+} channel)^{1,4}. Yet the role of all of these in sperm function has remained elusive, in part because it has not been possible to study sperm by patch-clamping, a central technique for investigating ion channels. Ren *et al.*² had similar difficulties with patch-clamping, but a variety of other experiments suggest that the channel they identified — which they dub 'CatSper' — is probably a Ca^{2+} -specific cation channel, and is certainly needed for normal sperm motility.

CatSper is the prototype of a new ion-channel family, described by the authors and by members of my laboratory in another paper⁵. The proteins in this family are something of an oddity. Channels such as the voltage-gated K^{+} channels consist of a single subunit, or 'repeat', which comprises six membrane-spanning portions and has a voltage sensor and an ion-selectivity pore. The common voltage-gated Na^{+} and Ca^{2+} channels consist of four such repeats. The CatSper^{2,5}, by contrast, have a single repeat, but the ion-selectivity pore is similar to that in each repeat of the voltage-gated Ca^{2+} channels. CatSper probably forms part or all of a tetrameric cation channel². Unfortunately, however, the ion selectivity of CatSper remains formally unproven: experimental expression of the protein alone or with other channel subunits resulted in no detectable ion-channel activity^{2,5}.

Nevertheless, the fact that CatSper is expressed only in male germ cells — specifically, in the tails of mature sperm — was a strong hint that it is involved in regulating sperm motility. Indeed, Ren *et al.* show that CatSper is required for normal progressive motility, and that its absence renders mice infertile. This represents a step towards understanding how ion channels regulate sperm motility. It also provides an opportunity to test the role of different forms of motility in fertilization. Sperm from CatSper-deficient mice swim with a progressive velocity about one-third that of normal. They can fertilize eggs whose extracellular matrix has been removed but not those with an intact matrix, so it seems that the reduction in progressive motility is sufficient to block penetration of the zona pellucida. Alternatively, the sperm might also fail to acquire the hyperactivated form of motility (a possibility that has not yet been tested).

The molecular details of how CatSper works remain unknown. Animals lacking CatSper produce normal quantities of morphologically normal sperm, so it probably

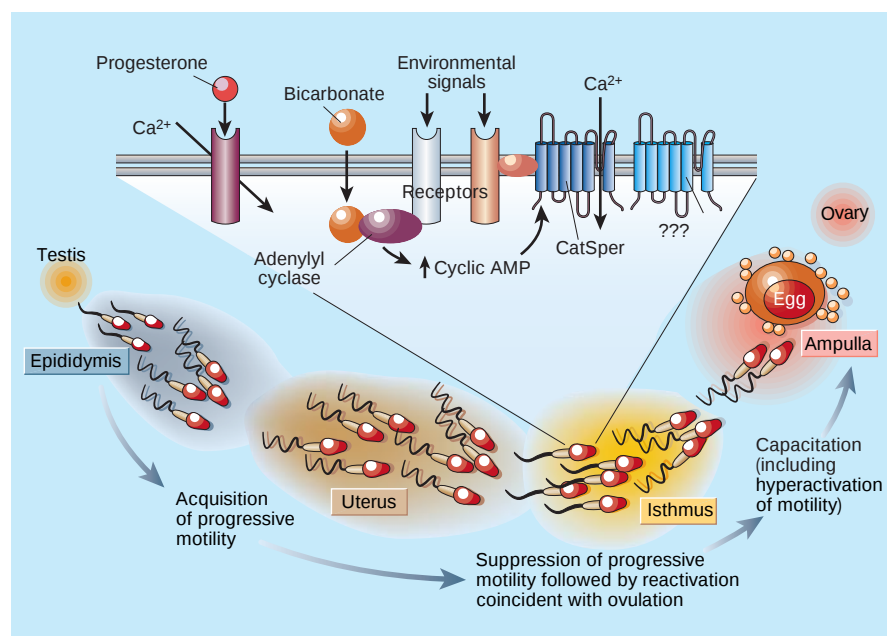


Figure 1 Changes in sperm motility, from production to fertilization. Spermatozoa are produced in the testis and acquire the ability to swim progressively (forwards) in the epididymis. In many species, after entering the female the speed of progressive motility is slowed in the isthmus of the oviduct. Later, often when ovulation occurs, sperm regain active motility and move to the ampulla, where fertilization occurs. While in the female, sperm cells gain the ability to fertilize the egg — a process known as capacitation. Part of this process is the induction of a whiplash form of motility (hyperactivated motility). Ren *et al.*'s results² suggest that a newly discovered ion channel, CatSper, is essential for sperm motility. Top, CatSper, in combination with partner subunits that have yet to be identified, might be a Ca^{2+} channel that is positively regulated by cyclic AMP. Concentrations of cyclic AMP in sperm are regulated by direct stimulation of the enzyme adenylyl cyclase by bicarbonate (HCO_3^-), or by environmental factors that act through an adenylyl-cyclase-coupled receptor. CatSper might also be directly regulated by a receptor that responds to environmental signals. Progesterone, possibly through a receptor similar to that for the neurotransmitter GABA¹¹, also causes an influx of Ca^{2+} , but this is independent of CatSper².

does not affect sperm development. But one could envisage that this channel is a central gateway through which different signalling pathways feed in information from the environment to regulate motility (Fig. 1). For example, sea-urchin sperm have a cell-surface receptor protein that binds to the egg's extracellular matrix⁶; activating this receptor leads to Ca^{2+} influx and induces the acrosome reaction. The receptor on sea-urchin sperm is similar to a human protein, PKD1, which might also be a receptor. (Mutations in PKD1 are associated with the kidney disorder autosomal dominant polycystic kidney disease.) PKD1 has been proposed⁷ to interact with and regulate an ion channel, PKD2. So it is plausible that a protein like the sea-urchin receptor could similarly interact with CatSper.

As mentioned above, cyclic nucleotides, such as cyclic AMP, have been reported to activate sperm motility. Ren *et al.*'s results² point to the idea that these molecules, too, might work through CatSper: sperm lacking CatSper did not respond to analogues of cyclic nucleotides by increasing their Ca^{2+} influx as usual. Cyclic AMP is generated by the enzyme adenylyl cyclase, which in sperm is connected to the environment through

bicarbonate⁸ or receptors that detect extracellular signals. Bicarbonate, present in genital fluids, has long been known to alter sperm function in mammals⁹, and signals from the egg that increase cyclic-nucleotide levels and sperm motility are also well known, at least in invertebrates¹⁰. Cyclic nucleotides might bind CatSper to affect its opening and closing directly, or they might be needed by enzymes that phosphorylate CatSper. The lack of a consensus cyclic-nucleotide-binding motif in CatSper supports the phosphorylation idea, but other data do not², so it remains unclear how cyclic nucleotides control CatSper.

Progesterone has also been reported to activate sperm motility, so it will be interesting to see what effect both progesterone and cyclic nucleotides have on the motility of sperm from CatSper-deficient mice. For example, if these molecules accelerate progressive motility, will fertility be restored? Or will that depend on the induction of hyperactivated motility? The results may show that the regulation of CatSper is critical to fertilization. This ion channel has a human counterpart², so it is already an attractive target for new contraceptives. It will become even more so if it proves to be



100 YEARS AGO

I should say a university is a place of higher education for those who are qualified by nature to profit by it. And I say that deliberately, holding the opinion as I do that it is not advisable to give more than elementary education to everybody, nor to encourage young people indiscriminately to enter upon a university course. An enormous amount of educational power is now wasted in trying to give a training to intellectual faculties which do not exist, for Providence has not given brains equally to everyone, and many a boy and girl now forced by parents or circumstances to the study of books would be much happier and more useful members of the community if they were taught to lay bricks and to sew and cook and wash, and do these necessary things well which are now done badly. This, of course, is not the business of a university, but if the university can so arrange its tests, whether by examination, or in some other way yet to be devised, as to prevent any large number of weaklings from entering upon the university curriculum, it will be doing a kindness to the rejected and a service to the rest of the world.

From *Nature* 10 October 1901.

50 YEARS AGO

When physics is regarded as the study of the external world, and the world is regarded as a distribution of matter in time and space, there is little room left for anything else except illusions...When I am asked whether I believe in the existence of the external world, I promise to reply when I am told precisely what is meant by the external world and its existence. I have not yet been called upon to redeem the promise. The fundamental question is: What exactly is it that physicists are doing? That can be answered satisfactorily only in terms of experience, not of the external world. It is in any event undeniable that all we can know of any world at all must come from our experience of it—not necessarily sense experience alone, but experience taken as a whole. If then we describe the task of science in terms of experience we shall describe it more directly than is possible in terms of an investigation of a postulated world. That might have been realized long ago; the physics of the past century has made its realization inescapable.

From *Nature* 13 October 1951.

a central node for processing extracellular information in sperm cells.

David L. Garbers is at the Cecil H. & Ida Green Center for Reproductive Biology Sciences, Department of Pharmacology, Howard Hughes Medical Institute, and the University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9051, USA.
e-mail: david.garbers@utsouthwestern.edu

1. Wennemuth, G., Westenbroek, R. E., Xu, T., Hille, B. & Babcock, D. F. *J. Biol. Chem.* **275**, 21210–21217 (2000).
2. Ren, D. *et al. Nature* **413**, 603–609 (2001).
3. Suarez, S. S. *J. Androl.* **17**, 331–335 (1996).

4. Jungnickel, M. K., Marrero, H., Birnbaumer, L., Lemos, J. R. & Florman, H. M. *Nature Cell Biol.* **3**, 499–502 (2001).
5. Quill, T. A., Ren, D., Clapham, D. E. & Garbers, D. L. *Proc. Natl Acad. Sci. USA* (in the press).
6. Mengerink, K. J., Moy, G. W. & Vacquier, V. D. *Zygote* **8**, S28–S30 (2000).
7. Emmens, S. W. & Somlo, S. *Nature* **401**, 339–340 (1999).
8. Zippin, J. H., Levin, L. R. & Buck, J. *Trends Endocrinol. Metab.* **12**, 366–370 (2001).
9. Okamura, N., Tajima, Y. & Sugita, Y. in *New Horizons in Sperm Cell Research* (ed. Mohri, H.) 197–203 (Japan. Sci. Soc. Press, Tokyo, 1987).
10. Garbers, D. L. *Annu. Rev. Biochem.* **58**, 719–742 (1989).
11. Serres, P. C. & Jouannet, P. *Biol. Cell* **92**, 255–266 (2000).

Materials science

Melting from within

Robert W. Cahn

For many years there has been uncertainty about the processes that trigger melting in solids. A new simulation manages to tie several threads together.

Melting has been a playground for theorists for almost a century. The problem is in understanding how and why a crystalline solid melts, and what determines the temperature at which this happens. Many theoretical criteria for melting have been advanced, of which two stand out: the Lindemann criterion¹ (1910) and the Born criterion² (1939). A paper by Jin *et al.*³ in *Physical Review Letters* has now addressed the question of whether the Lindemann and Born criteria predict the same melting temperature for an idealized crystal without surfaces. The new study, an international collaboration between China, Germany and the United States, is a highly sophisticated molecular-dynamics simulation, and shows that, as the crystal is 'heated', melting is triggered by instabilities governed simultaneously by the Lindemann and Born criteria.

Lindemann proposed that melting is caused by a vibrational instability in the crystal lattice when the root-mean-square displacement of the atoms reaches a critical fraction (δ_L) of the distance between them. Lindemann originally conceived δ_L as applying to the interior of the crystal, but in a later version it was applied to events at the surface, where the amplitude of atomic vibrations is larger than in the interior. Born, on the other hand, proposed that a 'rigidity catastrophe' occurs — caused by a vanishing elastic shear modulus — that determines the melting temperature within the bulk crystal. In other words, the crystal no longer has sufficient rigidity to withstand melting, so this process is often referred to as 'mechanical melting'. These two distinct theories have each accumulated an extensive literature.

It has been established experimentally that melting begins preferentially at a surface, and that superheating a crystal beyond the melting point set by the surface melting requires

this process to be impeded. For example, coating the surface with a metallic layer that has a higher melting point can suppress surface melting⁴ and retain the solid phase to bulk temperatures well above the equilibrium melting point, T_E . (This is the temperature at which both solid and liquid forms of a material can exist in thermodynamic equilibrium.)

Jin *et al.*³ set out to discover whether the Lindemann and Born criteria yield the same answer when both are applied to the bulk instead of a surface — that is, in the case of superheating. The authors considered a molecular-dynamics simulation of an array of about 7,000 particles obeying Lennard–

Jones interatomic forces, arranged on a face-centred cubic lattice, and heated using a step-by-step procedure. They also applied a set of three-dimensional boundary conditions to the simulation, which have the effect of making the surface of the crystal disappear. (This technique was pioneered in this context by Phillpot *et al.*⁵, who showed that, in the absence of an effective surface, a simulated silicon crystal can be superheated by hundreds of kelvin to a point at which mechanical melting occurs.)

In their simulation, Jin *et al.* first examined the validity of the Born criterion. They found that the idealized crystal could be superheated by about 20% (in terms of absolute temperatures); melting occurred when the elastic shear modulus of the crystal lattice came very close to zero. Mechanical melting was identified by a sudden change in the atomic volume.

Jin *et al.* then checked the Lindemann criterion. At the normal equilibrium melting temperature, T_E , the Lindemann parameter δ_L averages 0.12–0.13, which is consistent with Lindemann's original value for melting in the crystal interior; but at the much higher temperature T_M , required for mechanical melting, $\delta_L \approx 0.22$. This value is almost 80% larger than the value at T_E , but is similar to previous values found for surface melting at the equilibrium melting point. So $\delta_L \approx 0.22$ clearly has physical significance in both bulk and surface melting.

The last stage of the analysis is the most original. As the authors point out, the ability of the molecular-dynamics procedure to track "the physical properties of the atoms not only as global averages but also locally"

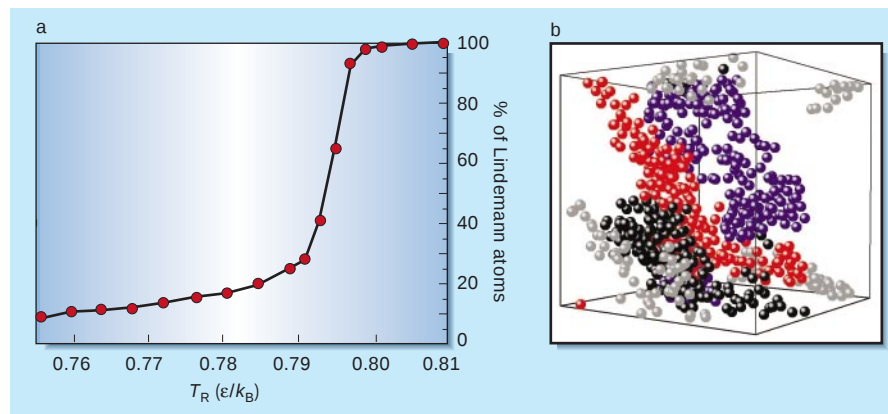


Figure 1 Heating without melting. a, The computed percentage of 'Lindemann atoms' in a crystal as a function of reduced temperature, T_R . A Lindemann atom is one for which $\delta_L > 0.22$, where δ_L is the root-mean-square displacement of the atom divided by the interatomic distance. In their simulation, Jin *et al.*³ show that the maximum temperature at which a crystal can be superheated (beyond its usual melting point, T_E) is $T_M = 0.79$. This is almost 20% higher than the normal equilibrium melting point ($T_E = 0.66$). (Reduced temperatures are given in units of ϵ/k_B , where ϵ is the depth of the potential well of the Lennard–Jones particles in the molecular-dynamics simulation, and k_B is the Boltzmann constant.) b, Snapshot of the positions of the Lindemann atoms in a three-dimensional idealized crystal without surfaces at $T_M = 0.79$. If the distance between a pair of Lindemann atoms is smaller than a critical value then they belong to the same cluster. In this region there are three large clusters with 219 (red), 214 (purple) and 187 (black) atoms. Jin *et al.*³ suggest that these clusters play a similar role to crystal surfaces in initiating melting.

is indispensable for exploring the correlation between the theories, and for investigating what happens on the atomic level at the onset of melting. To exploit this capability, Jin *et al.*³ define a 'Lindemann atom' as one for which the fractional root-mean-square displacement exceeds the critical value $\delta_L \approx 0.22$. The number and location of these Lindemann atoms was examined as a function of temperature, and they were found to increase rapidly in number near the melting temperature (Fig. 1a) and to form clusters of increasing size as the crystal heated up (Fig. 1b).

The authors then calculated the Born elastic moduli for the Lindemann atoms, and found that the average value of the elastic shear modulus is much closer to zero for the totality of these particles than for the crystal as a whole. To quote the carefully worded conclusion of the paper: "These results demonstrate a strong correlation between the Lindemann criterion and the Born criterion: melting is initiated by local lattice instabilities governed by both." As they see it, the simultaneous violation of both the Linde-

mann and Born criteria "sets the stage for the emergence of local regions that play the role of surfaces in initiating the equilibrium melting". The temperature at which these processes take place represents the maximum feasible temperature for superheating.

Perhaps the most surprising conclusion of this paper is that, at the superheating limit T_M , the clusters have properties similar to the surface atoms at the much lower equilibrium melting temperature, T_E . This is a major discovery that may well finally settle the long-standing uncertainty about the correct criterion for melting. It is also an example of what computer simulation can achieve when all of its capabilities are properly exploited. ■

Robert W. Cahn is in the Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK.

e-mail: rwc12@cam.ac.uk

1. Lindemann, F. A. *Z. Phys.* **11**, 609 (1910).
2. Born, M. *J. Chem. Phys.* **7**, 591 (1939).
3. Jin, Z.-H., Gumbsch, P., Lu, K. & Ma, E. *Phys. Rev. Lett.* **87**, 055703 (2001).
4. Cahn, R. W. *Nature* **323**, 668–669 (1986).
5. Phillpot, S. R., Yip, S. & Wolf, D. *Computers Phys.* **3**, 20–31 (1989).

Gene expression

The odd coupling

R. Andrew Keys and Michael R. Green

Gene expression requires several complex biochemical reactions to be carried out precisely. These reactions must also be coordinated, and molecular biologists now have a better handle on how that happens.

In organisms whose cells have a nucleus, the expression of a protein-encoding gene involves several steps: transcription of the gene; extensive processing of the initial product, the 'precursor' messenger RNA; and export of the correctly processed 'mature' mRNA from the nucleus to the body of the cell. A great deal has been learned about the mechanisms and factors involved through biochemical and genetic studies of each reaction. The challenge now is to understand how these steps fit together correctly. Writing on pages 644 and 648 of this issue, Luo and colleagues¹ and Sträßer and Hurt² provide insight into how two reactions — the 'splicing' of precursor mRNAs and the export of processed mRNAs — may be coupled.

When a nuclear gene is transcribed, the precursor mRNA (pre-mRNA) product consists of protein-encoding portions (exons) that are interspersed with non-coding sequences (introns). During splicing, the introns are excised and the exons joined together. It has long been suspected that splicing and export must in some way be connected, to ensure that only the final mRNA product — and not unspliced pre-mRNA, splicing intermediates or excised introns — leaves the nucleus. Consistent

with this suspicion, mRNAs are known to be retained in the nucleus as a result of being assembled into the spliceosome, the complex that carries out splicing. After splicing, the mature mRNA is released from its spliceosomal bondage and is free to be exported.

Studies of multicellular organisms have shown that splicing actually enhances the export of mRNA. The increased efficiency of export results from assembly of the mRNA into a special RNA–protein complex, which differs from that assembled on an otherwise equivalent RNA that has not undergone splicing³. This complex contains a protein known as Aly^{4,5} (also called REF⁶ and BEF⁷), the yeast counterpart of which, Yra1, is known to be required for mRNA export⁸. According to the current model, Aly/Yra1 mediates mRNA export through interaction with another protein, called TAP in mammals and Mex67 in yeast. TAP/Mex67 in turn interacts with the nuclear pore complex^{6,8}, the structure through which molecules move into and out of the nucleus. Related studies have shown that splicing results in deposition of at least five proteins on mature mRNA near the junction between exons. The resulting 'exon-junction complex' contains both Aly and TAP, as well as proteins that

are thought to be involved in other mRNA-related functions⁹.

The specific issue addressed in the new papers^{1,2} is the mechanism by which Aly/Yra1 is recruited to spliced mRNA. Luo *et al.*¹ carried out a series of experiments involving injecting specific proteins into frog eggs, and found that Aly and UAP56, an essential human splicing factor¹⁰, interact both physically and functionally. In particular, injection of excess UAP56 blocked the export of mature mRNA by sequestering Aly, preventing it from binding to mRNA.

This was completely unexpected — there was no reason to suspect that Aly and UAP56 would pair up. UAP56 is a member of a family of proteins known as DEXD/H-box proteins, which mediate the ATP-dependent events that drive splicing. Several of these proteins can unwind RNA, an activity that is

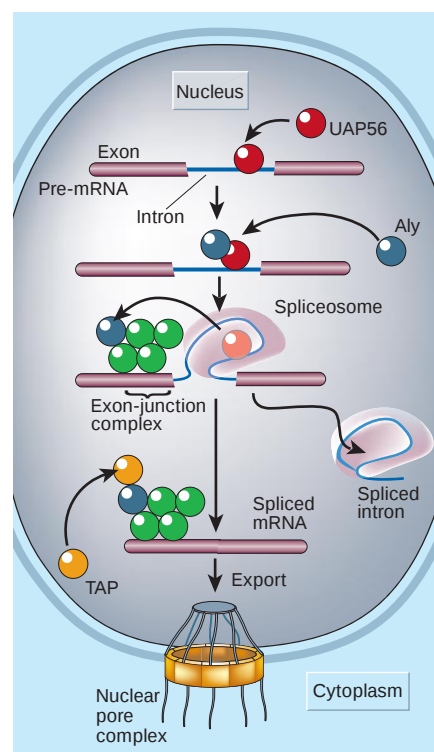


Figure 1 A model to explain coupling between the splicing of precursor messenger RNAs (pre-mRNAs) and the export of the mature mRNA product from the nucleus. Luo *et al.*¹ and Sträßer and Hurt² have found that the mammalian protein Aly (or its yeast counterpart, Yra1; not shown) interacts with the splicing factor UAP56 (Sub2 in yeast). These and previous results suggest a model in which first UAP56 and then Aly is recruited to pre-mRNA (top). Next, the spliceosome — a large RNA–protein complex — is assembled on the pre-mRNA and carries out splicing (removal of the intron and joining together of the two exons). At some point during this process, Aly is moved to the exon-junction protein complex. The mRNA-export factor TAP then binds to Aly and, in so doing, targets the mature mRNA to the nuclear pore complex and transports it out of the nucleus.

thought to promote the structural changes in the spliceosome that are required for splicing to occur¹¹. Human UAP56 is needed for the first ATP-dependent step in splicing — the interaction between one of the spliceosomal RNAs and the 'branch-point' sequence in the intron¹⁰.

Sträßer and Hurt², meanwhile, carried out a genetic screen in yeast to identify proteins that interact with Yra1. The screen revealed that Sub2 (also called yUAP), the yeast counterpart of human UAP56 (refs 12–14), genetically interacts with Yra1. Having fingered Sub2 as a potential export factor, the authors showed that mutation or over-expression of this protein results in a general defect in mRNA export. Finally, they found that Sub2 binds to Yra1, and that the binding of Sub2 and Mex67 to Yra1 are mutually exclusive. Together with previous studies, the new results^{1,2} suggest a model in which Aly/Yra1 is first recruited to pre-mRNA through interaction with the splicing factor UAP56/Sub2. This factor is then displaced by TAP/Mex67, which targets the spliced mRNA to the nuclear pore complex (Fig. 1).

Not surprisingly, these results raise new questions. Perhaps most unexpected is Sträßer and Hurt's finding² that Sub2, a bona fide splicing factor^{12–14}, is required for the export of both spliced mRNAs and RNAs that never contained introns. The involvement of a splicing factor in the export of an intronless RNA is not without precedent¹⁵. Nonetheless, one is prompted to ask how Sub2 is directed to RNAs that lack splicing signals. And are the ATP-hydrolysing and RNA-helicase capabilities of Sub2 needed in processing such RNAs?

A second question is whether there are other proteins with functions similar to that of UAP56/Sub2. Although Sub2 is normally essential for viability, it is dispensable in certain mutant yeast strains¹². But Sträßer and Hurt² found that nuclear export is at most only modestly compromised under these circumstances. This observation is most readily explained by the existence of a protein, yet to be identified, that can substitute for Sub2 in export.

Third, UAP56/Sub2 is thought to function at the branch-point sequence in introns^{10,12–14}, so this is where one would expect Aly/Yra1 to be recruited. But on spliced mRNA, Aly/Yra1 is found in the exon-junction complex⁹, raising the question of how Aly/Yra1 moves from intron to exon.

Finally, Aly was originally identified as a protein that is involved in transcription^{5,7}. Apparently, then, it participates in several steps in gene expression. How can one protein be involved in reactions as seemingly diverse as transcription and mRNA export? A clue may be that Aly has properties that are characteristic of molecular chaperones⁹, a class of proteins that help other proteins to fold into the correct shape. A parsimonious

(and testable) theory is that Aly functions as a chaperone in several nuclear events that have in common the assembly, remodelling or disassembly of multiprotein complexes on nucleic-acid scaffolds. Transcription and splicing factors could function as 'adaptor' proteins that recruit a single factor, such as Aly, to different substrates to participate in seemingly unrelated events. If this proves to be the case, the integration of nuclear processes may be much more extensive than was originally imagined. ■

R. Andrew Keys and Michael R. Green are at the Howard Hughes Medical Institute and the Programs in Gene Function and Expression, and Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA.

Nanotechnology

Molecules join the assembly line

Paul S. Weiss

Making patterns from molecular building blocks sounds like child's play, but has been surprisingly difficult to do. A new approach to assembling molecules into patterns may ultimately lead to molecule-based devices.

Synthetic molecules can be designed to form assemblies with predictable structures, rather like the complex molecular assemblies found in nature. Such artificial assemblies are seen as a means of bridging the gap between the small molecules created by synthetic chemistry and the larger structures made using nanolithography that can be linked to the outside world. Chemists can already tailor molecules that have specific functions with atomic-scale precision; the objective now is to develop technologies that exploit them. But one of the biggest challenges is bridging the gap between molecular features 1 nanometre across and device features on the order of 100 nanometres. Ultimately, the two will have to meet in the middle.

On page 619 of this issue, Yokoyama *et al.*¹ describe a step from the bottom up — they

e-mail: michael.green@umassmed.edu

1. Luo, M.-J. *et al.* *Nature* **413**, 644–647 (2001).
2. Sträßer, K. & Hurt, E. *Nature* **413**, 648–652 (2001).
3. Luo, M. & Reed, R. *Proc. Natl Acad. Sci. USA* **96**, 14937–14942 (1999).
4. Zhou, Z. *et al.* *Nature* **407**, 401–405 (2000).
5. Bruhn, L., Munnerlyn, A. & Grosschedl, R. *Genes Dev.* **11**, 640–653 (1997).
6. Stutz, F. *et al.* *RNA* **6**, 638–650 (2000).
7. Virbasius, C. M., Wagner, S. & Green, M. R. *Mol. Cell* **4**, 219–228 (1999).
8. Sträßer, K. & Hurt, E. C. *EMBO J.* **19**, 410–420 (2000).
9. Lykke-Andersen, J. *Curr. Biol.* **11**, R88–R91 (2001).
10. Fleckner, J., Zhang, M., Valcarcel, J. & Green, M. R. *Genes Dev.* **11**, 1864–1872 (1997).
11. Staley, J. P. & Guthrie, C. *Cell* **92**, 315–326 (1998).
12. Kistler, A. L. & Guthrie, C. *Genes Dev.* **15**, 42–49 (2001).
13. Libri, D., Graziani, N., Saguez, C. & Boulay, J. *Genes Dev.* **15**, 36–41 (2001).
14. Zhang, M. & Green, M. R. *Genes Dev.* **15**, 30–35 (2001).
15. Huang, Y. & Steitz, J. A. *Mol. Cell* **7**, 899–905 (2001).

have designed molecules with prescribed attachments that fit together in specific ways, much like a molecular tinker toy set. They use the exquisite control provided by molecular synthesis to govern the formation of molecular assemblies over scales as large as 100 nm.

Nature is replete with examples of how to join molecules together to obtain structures on many length scales. Most protein structures are formed from strongly bound repeating units called polypeptide chains, which assemble into structures governed by weaker hydrogen bonds and van der Waals interactions. These include regular structures such as sheets, helices and what appear to be walls made of miniature bricks. Chemists and others have borrowed molecules from biology, most notably DNA, to make structures of their own. For example, Seeman and his group² have constructed three-dimensional

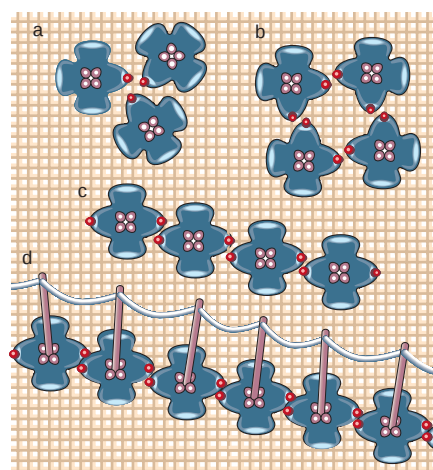


Figure 1 Patterns with porphyrins. Yokoyama *et al.*¹ show that, by selecting the number of linkages (red) between porphyrin molecules and the orientation of these linkages with respect to one another, both localized and extended structures can be formed. a, With only one linkage per molecule, localized assemblies of three molecules form. b, With two linkages per molecule and these linkages at 90° relative to one another, localized assemblies of four molecules form. c, With two linkages per molecule at 180° relative to one another, extended linear structures form. d, By taking advantage of the chemistry of the centre of the molecule, these extended structures could be used to guide the formation of superstructures.

objects and two-dimensional tapes supported on substrates using carefully selected combinations of DNA oligonucleotide sequences. Kawai and co-workers³ have looked at the interactions of the DNA bases themselves and the structures they form when constrained to two dimensions on substrates. With device applications in mind, Braun *et al.*⁴ have used complementary DNA strands to link metal pads together electrically by using DNA as a template for metal deposition once the desired links are formed.

Synthetic systems have been less amenable to creating precise assemblies. Yokoyama *et al.*¹ developed a simple set of rules for making extended molecular assemblies — in this case, from porphyrin molecules adsorbed on a gold surface (Fig. 1). By working on a surface, they require the molecules to lie flat. Each porphyrin molecule has four potential points of attachment located at its periphery, as if at the four points of a compass. By tailoring the molecules to form one or more of these attachments, the authors can control the number of connections and the geometric relations with neighbouring molecules. Specifically, they replace one or more of the four 'compass points' with a cyanophenyl group, which interacts in a pre-determined way with other cyanophenyl groups.

When there is only one cyanophenyl connection, the porphyrin molecules form three-molecule rings linked together in the centre (Fig. 1a). If instead there are two connections separated by an angle of 90°, small clusters of four molecules are formed (Fig. 1b). With two connections located at opposite ends of the molecule, long chains form (Fig. 1c), which can be up to 100 nm long. In every case, the modified porphyrin molecules find one another on the surface and link together spontaneously; they do not need to be patterned on the surface in any other way.

The ability to guide the formation of such structures offers several benefits. In polymer chemistry, a well-known approach for determining the final properties of the material involves varying the number of connections and the mix of component molecules. For example, the number of crosslinks between polymer chains (and therefore the final rigidity of the material) is controlled by the fraction of molecules that can form more than two linkages. In Yokoyama and colleagues' study, the synthesized molecules are selected to control both the degree of connectivity and the spatial arrangement of the molecular assemblies. In the case of the porphyrin assemblies the connections between molecules are likely to be dipole-dipole interactions, rather than the covalent bonds formed in crosslinked polymers.

The next steps are to link these molecular assemblies to the outside world through microscopic structures created using standard nanofabrication techniques⁴, and to

add function to the assembled molecules. The molecules chosen by Yokoyama *et al.* are particularly suited to this latter purpose because porphyrins can bind metal atoms at their centres — the metals can be chosen to link further molecules across the top of the assembly, so as to form an extended superstructure (Fig. 1d). This approach could enable researchers to separate the design of chemical linkages required for assembly from those that would be a part of the desired function. In electronic devices based on single molecules, for example, the chemical contact with the outside world (usually through electrodes) can dominate the electrical resistance of the entire system^{5–7}. By separating structural constraints from molecular function, there is more room to optimize the functional part of the assembly independently.

Much work remains to make such functional assemblies a reality. Practical assem-

blies will need to be able to cope with surface defects, such as steps, and with useful substrates: semiconductors rather than gold, for example. The strategy outlined by Yokoyama *et al.* may be flexible enough both to build nanofabricated contacts and to provide the platform on which to construct and align more complex structures. ■

Paul S. Weiss is in the Department of Chemistry, Pennsylvania State University, University Park, Pennsylvania 16802-6300, USA.

e-mail: stm@psu.edu

1. Yokoyama, T., Yokoyama, S., Kamikado, T., Okuno, Y. & Mashiko, S. *Nature* **413**, 619–621 (2001).
2. Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. *Nature* **394**, 539–544 (1998).
3. Furukawa, M., Tanaka, H. & Kawai, T. *J. Chem. Phys.* **115**, 3419–3423 (2001).
4. Braun, E., Eichen, Y., Sivan, U. & Ben-Yoseph, G. *Nature* **391**, 775–778 (1998).
5. Joachim, C. *New J. Chem.* **15**, 223–229 (1991).
6. Bumm, L. A., Arnold, J. J., Dunbar, T. D., Allara, D. L. & Weiss, P. S. *J. Phys. Chem. B* **103**, 8122–8127 (1999).
7. Yao, Y., Kane, C. L. & Dekker, C. *Phys. Rev. Lett.* **84**, 2941–2944 (2000).

Cloud physics

Inside history on droplets

Marcia Baker

Upper-tropospheric clouds contain ice particles, most of which result from the freezing of liquid droplets. That freezing, it emerges, is far more complicated than had been thought.

Clouds in the troposphere, the atmospheric layer between the ground and the stratosphere, have profound influences on air chemistry, weather and global climate. Many of the particles in these clouds originate as small droplets that grow by condensation of water vapour as the droplets rise in the atmosphere. Some of the ice that forms in the clouds originates in the clear air outside the droplets. But most of it forms as the droplets freeze during their ascent, and predicting the conditions under which droplet freezing occurs is one of the main goals of cloud physicists. This is the background to a laboratory study by Zuberi and colleagues¹, published in *Journal of Physical Chemistry A*. The paper highlights the potential significance of droplet history on droplet freezing, and points to new directions for research.

In the atmosphere, ice is the thermodynamically stable state of water below 0 °C. However, water can remain liquid, in a metastable ('supercooled') state, down to much lower temperatures. Pure water droplets do not freeze spontaneously at temperatures above about –37 °C, and the freezing temperatures of droplets of aqueous solution are even lower. In the absence of external surfaces, a volume of water or aqueous solution is homogeneous, and freezing is equally likely to occur at any point within it — freezing under these conditions is therefore called homogeneous. In contrast, if a solid surface

is immersed in or in contact with the droplet, liquid freezing can be catalysed by that surface at relatively high temperatures; this is called heterogeneous freezing².

Zuberi *et al.* explored the freezing behaviour of droplets consisting of concentrated solutions of soluble salts that are commonly found in the atmosphere. By subjecting the droplets to a range of temperatures, varying over time, they caused the salts to crystallize inside the droplets and then to redissolve (partially or completely), after which the droplet temperatures were lowered until freezing occurred. The authors found that the morphology and sizes of the salt crystallites depended on the thermal histories of the droplets, as did the temperatures at which the droplets eventually froze. Freezing was homogeneous in some cases but was clearly heterogeneous in others.

The effects of droplet history have not been understood previously and may explain the differences between the results of some other experiments³ and those of Zuberi *et al.* The new findings cast doubt on the idea that heterogeneous — high-temperature — freezing of a droplet can occur only when it comes into contact with pre-existing solid surfaces.

What do these findings tell us about events in the troposphere? Freezing patterns in atmospheric clouds show tremendous variability, with wide ranges of observed freezing temperatures that we often cannot



Figure 1 Clouds in the upper troposphere: their climatic effects depend on the frozen water content.

predict accurately. Some observations show liquid droplets that persist until very low temperatures, suggesting that the freezing is homogeneous. In other studies, however, large quantities of ice form at relatively high temperatures, implying that heterogeneous freezing is taking place. This variability has prompted energetic speculation on the nature of 'ice nuclei' — pre-existing solid particles in clouds that are particularly efficient at catalysing freezing on their surfaces at relatively high temperatures.

Several studies have involved collecting cloud air and investigating the temperatures at which the particles in it can lead to droplet freezing. The most likely candidate nuclei include mineral dust and clays, particles of biological origin such as plant debris, and some products of industrial combustion. The experiments of Zuberi *et al.* suggest, however, that this approach, which is based on the idea that heterogeneous freezing of droplets requires pre-existing ice nuclei, may not always be sufficient to predict the eventual temperatures of droplet freezing. Rather, it seems that the formation of ice nuclei within droplets may sometimes be part of the freezing process itself.

There are, of course, many differences between conditions in laboratory experiments and those in the atmosphere, so the applicability of the results of these new experiments is likely to be limited. Nonetheless, as Zuberi *et al.* point out, there are times and places in the atmosphere at which analogous transformations of droplets occur, which might affect the freezing modes. Close to the boundaries between clouds and clear air, for example, aerosol particles and cloud droplets can be subjected to one or more cycles of condensation and evaporation, accompanied by large temperature fluctuations. If a droplet undergoes these processes at low temperatures it can freeze, and if it contains inorganic or organic materials of low solubility (such as some organics that are commonly found in atmospheric aerosol particles⁴), the freezing

mode and temperature could depend on the droplet's history.

Some cases⁵ of so-called 'ice multiplication' at temperatures above about -20°C might also be explained in this way. This phenomenon, in which ice-particle concentrations far exceed the concentrations of known ice nuclei, may sometimes be due to heterogeneous freezing onto the solid inclusions in originally liquid droplets. Ideas such as these need to be investigated under realistic atmospheric conditions. If it turns out that droplet history is important in determining freezing temperatures, then that history must be taken into account in modelling and laboratory studies of tropospheric clouds.

The effects of upper-tropospheric clouds on climate are highly sensitive to the amount of frozen cloud water. For instance, according to the latest report from the Intergovernmental Panel on Climate Change⁶, differences in assumptions about how much water is frozen lead to differences of up to 17 W m^{-2} in the global average flux of radiation entering or leaving Earth. For comparison, the change in flux due to the CO_2 increase over the past 200 years is less than 2 W m^{-2} . Droplet freezing is one of the central processes in cloud physics but also one of the least understood. We badly need new ideas in this area, and experiments such as those of Zuberi *et al.* are a great help in providing fresh approaches to the subject.

Marcia Baker is in the Departments of Atmospheric Science, and Earth and Space Sciences, PO Box 351650, University of Washington, Seattle, Washington 98195, USA.
e-mail: marcia@geophys.washington.edu

1. Zuberi, B., Bertram, A. K., Koop, T., Molina, L. T. & Molina, M. J. *J. Phys. Chem. A* **105**, 6458–6464 (2001).
2. Pruppacher, H. D. & Klett, J. *Microphysics of Clouds and Precipitation* (Kluwer, Dordrecht, 1997).
3. Chen, Y. L. *et al.* *J. Atmos. Sci.* **57**, 3752–3766 (2000).
4. Jacobson, M. C., Hansson, H. C., Noone, K. J. & Charlson, R. J. *Rev. Geophys.* **38**, 267–294 (2000).
5. Rangno, A. & Hobbs, P. V. *Q. J. R. Met. Soc.* **117**, 207 (1991).
6. Intergovernmental Panel on Climate Change *Climate Change 2001: The Scientific Basis* Ch. 5 (Cambridge Univ. Press, Cambridge, 2001); <http://www.ipcc.ch>

Daedalus

Even more light

Artificial light is one of the few great inventions: it frees mankind from dependence on sunlight. Artificial light often depends on a hot object — a filament or gas mantle — as close as possible to the temperature of the Sun. In practice, 3,000 K is good. Alternatively, fluorescent materials can be excited by ultraviolet light or electric discharges, as in gas-discharge tubes. But both types of light are deplorably inefficient.

Daedalus now recalls the free-electron laser. A beam of electrons is fired along a tube, and set wiggling by magnets lining it. Electrons sent along a wiggly wire as an ordinary current should wiggle just as much, and their sideways oscillation should emit visible light. Modern micro-deposition can probably impose 10^{12} wiggles per metre on a conductive pathway, so a current of 10^3 metres per second should emit blue light at a frequency of 10^{15} Hz. A high-temperature superconductor may be needed, but the thing looks feasible: an amp is some 6×10^{18} electrons a second.

DREADCO engineers are therefore depositing cuprate superconductors, and even conducting metals such as silver, onto transparent quartz plates. They are packing as many wiggles as possible into the conductor. Transformer coupling to the wiggly element should let ordinary mains electricity drive the new light. It should compete strongly with existing lamps. It will be far more efficient than they are. It will hardly get warm; indeed, it may need to be cooled, but Daedalus hopes that room-temperature superconduction will be possible soon.

The new 'electrolamp' will emit a very pure light, its frequency defined by the speed of the electrons and the size of the wiggles. The first samples will go at a vast price for visible, infrared and even ultraviolet spectrometers; these will recover the heavy development costs. When the lamp reaches commercial production, several parallel 'filaments' will each have different number of wiggles, so as to emit a graded rainbow of colours approximating to white (as with gas-discharge tubes today).

Photographers will love a special electrolamp, whose three filaments are each tuned to one of the dyes in the film. It should combine great sensitivity with accurate colour rendition. The broader, whiter electrolamp should transform the domestic scene. Equally effective on mains or battery, it will complete our victory over fitful solar illumination.

David Jones

Obituary

Edward T. Hall (1924–2001)

Edward Hall, who died on 11 August aged 77, was a physical scientist, an ingenious inventor and a skilled craftsman. He was without doubt a most unconventional man. He was wealthy, a discerning collector of Chinese pottery and historical clocks, and a generous benefactor; and in his retirement he built what is reputed to be the most accurate pendulum clock in the world. As to his scientific work, which was notable, he confessed that an important part of it was the sheer fun of doing it.

For the past 45 years, Hall's driving interest, at least in the technical part of his life, was the pursuit of archaeological science — the invention, construction and use of methods for finding, dating and authenticating ancient objects, and for throwing light on ancient technology. These years saw a worldwide growth of interest in this hybrid of science and the arts, now known as 'archaeometry', and that growth owes a great deal to Hall's energy and flair.

Hall's family prosperity came from his paternal grandfather, who settled in England after an exceedingly profitable gold-mining venture in Australia. Hall himself was born in London in May 1924. He was at school at Eton until 1943, served in the Royal Navy until 1946, and then entered New College, Oxford, to read for his degree in chemistry.

In Hall's postgraduate years at Oxford University, from 1949, his studies shifted from chemistry to physics, and, thanks to his personal wealth, he was free to ardently follow his interest in applying the methods of physics to art and archaeology. I believe that if he had not been able to do so within the university, he would still have

continued some of the work at his own expense in the laboratory and workshop he built at his home. But he obtained the backing of professors of both physics and archaeology to start a specialist laboratory for the purpose. Significantly, it was to have close links to both of those faculties but not be subject to either; anyone who has been close to academic politics and the supposed cultural divide can imagine the stormy passage faced by such a proposal. But in 1953 there surfaced the Research Laboratory for Archaeology and the History of Art, which Hall so effectively directed for the next 35 years.

His first aim, knowing the value of the museum objects he would be handling, had been to develop non-destructive methods of analysing them. He successfully built an X-ray fluorescence spectrometer, using at first an X-ray tube and automatic control equipment he had made himself, to determine the alloy composition of Greek coins. He then turned to the pigments and glazes on Chinese pottery borrowed from museums. To his horror he found that after the long X-ray exposure needed for a complete analysis, his 'non-destructive' method had left brown spots in the glaze of the first piece he examined. He promptly abandoned the test and rushed to the London sale-rooms to buy a replacement. However, this pioneer test soon paved the way for quicker measurements on a large array of Chinese blue-and-white ware, willingly lent by museums and private collectors, and drawn from his own private collection. The results threw light on the sources of pigments used by Chinese potters, and formed the first substantial

research report coming from Hall's Oxford laboratory.

Hall gained personal fame, and helpful publicity for the embryonic discipline of archaeometry, by his 1949 analysis of surface stains on the bones of the so-called Piltdown Man, for which his X-ray spectrometer was ideally suited. He thereby contributed to the exposure in 1953 of this anthropological misfit as a deliberate fraud.

He successfully developed other analytical instruments, including one of the earliest electron-beam microprobes. He pioneered a portable analytical instrument which could be taken directly to an object in a museum, and which became widely used. These, and other successful instruments developed by Hall and his colleagues, have been produced commercially for the benefit of archaeometrists throughout the world by the light engineering company into which Hall transformed his personal laboratory and workshop.

Hall was also the prime mover in designing and making several devices for locating buried archaeological remains. These included the use of the free precession of the proton to detect magnetic-field anomalies caused by those remains, and one of his special interests was the design of an underwater proton magnetometer, which he successfully used in searching for wrecks from his small motor cruiser off the Turkish coast.

More recently, Hall was closely involved with installing an accelerator mass spectrometer, dedicated to radiocarbon dating of organic material, in the Oxford laboratory. With this elaborate and costly equipment it is possible to date ancient specimens of very small mass. Some 10,000 samples have so far been dated by the laboratory, of which the most celebrated is without doubt a postage-stamp-sized piece of cloth cut from the Shroud of Turin. This was put to the test in 1988, and the result (a medieval date) brought Hall and his laboratory further into the public eye.

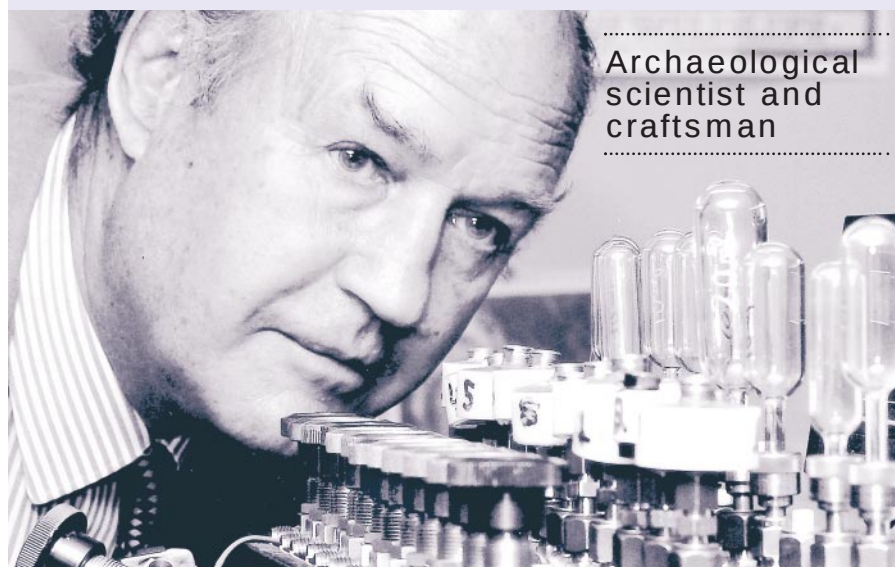
Hall was a friendly, generous and unpretentious man, bold in his decisions, and widely respected by a large circle of friends within and beyond the sciences. He was adviser or trustee to many of the national institutions that he could help in selecting and conserving antiquities. He found it hard to resist the appeal of any new and unusual exploit, which he would enter with great zest. Those among us who can remain steadily on track despite such diversions are rare birds. Teddy Hall was a very rare bird indeed.

Stuart A. Young

Stuart A. Young is at PO Box 4, Noordhoek 7985, South Africa.

e-mail: obuzzard@iafrica.com

Archaeological
scientist and
craftsman



Reward value of attractiveness and gaze

Making eye contact enhances the appeal of a pleasing face, irrespective of gender.

Faces are visual objects in our environment that provide strong social cues¹, with the eyes assuming particular importance^{2,3}. Here we show that the perceived attractiveness of an unfamiliar face increases brain activity in the ventral striatum of the viewer when meeting the person's eye, and decreases activity when eye gaze is directed away. Depending on the direction of gaze, attractiveness can thus activate dopaminergic regions that are strongly linked to reward prediction⁴, indicating that central reward systems may be engaged during the initiation of social interactions.

In an event-related functional magnetic resonance imaging (fMRI) study, 16 subjects (8 male, 8 female) were shown colour images of 40 different faces, which had their eyes directed either at or away from the subject (Fig. 1). After the fMRI scanning session, subjects rated the attractiveness of the faces they had seen, and these ratings were normalized and parametrically entered into the

analysis. Significant effects ($P < 0.05$, corrected for multiple comparisons) were assessed using a *t*-test and displayed as statistical parametric maps using SPM99b software.

No brain region showed any activation in response to facial attractiveness *per se*. We therefore investigated whether processing of attractiveness depends on gaze direction. When eye gaze was directed at the subjects, activity in the ventral striatum correlated positively with attractiveness (slope: +0.46% change in signal intensity per standard deviation increase in attractiveness rating). In contrast, this correlation was reversed when eye gaze was averted, causing activity in the ventral striatum to decrease with increasing attractiveness (slope: -0.88% signal intensity per s.d. increase in attractiveness).

Figure 2 shows the resulting activation map when these two contrasts are combined as an interaction — for example, the contrast (correlation with attractiveness when eye contact is made) minus (correlation with attractiveness when no eye contact is made). In a fixed-effects analysis, the effect was significant at a voxel level (the smallest volume tested in our study) in the right ventral striatum ($P = 0.015$) and at cluster level bilaterally (right, $P < 0.01$; left, $P = 0.01$). We also implemented a second-level random-effects analysis to determine whether the results could be extended to the population at large. Significant effects persisted at a cluster level in the right ventral striatum ($P < 0.01$; point of maximal activation: $x = 12$, $y = -8$, $z = 10$) and approached significance in the left ventral striatum ($P = 0.06$, $x = -10$, $y = 0$, $z = 4$).

The gender of the observed face in relation to the gender of the observer did not influence the response in the ventral striatum. This might indicate that features depicting attractiveness (such as facial form and symmetry) are low-level cues that are processed independently of gender.

Our results provide evidence not only that the ventral striatum processes reinforcing stimuli that serve basic survival needs⁵, but also that it is more generally involved in the evaluation of stimuli with relevance to social interactions. The activity of dopaminergic neurons that project to the ventral striatum is associated with reward prediction⁶. The firing rate of these neurons follows a specific pattern, increasing when an unexpected reward is seen and decreasing when an expected reward fails to materialize, thereby predicting error in future rewards.

In accordance with this specific pattern, we found no activation associated with

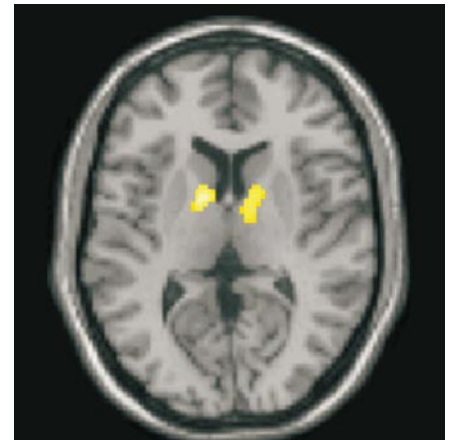


Figure 2 Depending on eye gaze, the degree of attractiveness of faces modulates activity in the ventral striatum, a brain area associated with reward prediction. When eye gaze is directed at a subject, the degree of attractiveness of individual faces correlates positively with brain activity in the ventral striatum; when eye gaze is directed away, this correlation is reversed and activity in the ventral striatum decreases with increasing attractiveness. Activation maps are superimposed on the surface of a T1-weighted template (fixed-effects analysis showing areas in which activity surpasses the threshold of $P = 0.05$ at the cluster level. Time to repetition, 3.16 s; 32 transverse slices; voxel size, $3 \times 3 \times 3$ mm; Siemens VISION 2T).

attractiveness *per se*, but only as an interaction with eye gaze. Assuming that the dimension of attractiveness constitutes a reward, especially when social interaction is initiated, returned eye gaze from an attractive face represents a more favourable result than expected, leading to enhanced responses, whereas failing to make eye contact with an attractive face is a disappointing outcome, leading to reduced activity in the dopaminergic target systems. Likewise, eye gaze from an unattractive face may result in disappointment and a reduced response, whereas missing eye contact with an unattractive face may be a relief and thus enhance activity. The profile of the response we describe here may reflect an automatic evaluation of the likely reward that can be derived from conspecifics.

Knut K. W. Kampe*, **Chris D. Frith†**,
Raymond J. Dolan†, **Uta Frith***

**Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK*

e-mail: k.kampe@ucl.ac.uk

†Wellcome Department of Cognitive Neurology, 12 Queen Square, London WC1N 3BG, UK

1. Bruce, V. & Young, A. Br. J. Psychol. **77**, 305–327 (1986).
2. Kleinke, C. L. Psychol. Bull. **100**, 78–100 (1986).
3. Emery, N. J. Neurosci. Biobehav. Rev. **24**, 581–604 (2000).
4. Schultz, W., Dayan, P. & Montague, P. R. Science **275**, 1593–1599 (1997).
5. Berridge, K. C. Neurosci. Biobehav. Rev. **20**, 1–25 (1996).
6. Schultz, W. J. Neurophysiol. **80**, 1–27 (1998).



Figure 1 Examples of eye-contact and non-eye-contact stimuli. Faces were counterbalanced for head position, sex and, for non-eye-contact images, gaze direction (right or left). During each scanning session, subjects were presented with 2 runs of 80 faces and 40 scrambled control faces in a randomized, event-related design (new image every 3.5 s; presentation time, 1.2 s, with intervening scrambled images). No task was required from subjects while they viewed these images, but to secure and test their attention, six target images (faces with closed eyes) were included, to which subjects were asked to respond by pressing a button. In statistical analysis, these latter images were modelled as events of no interest and excluded from further evaluation. After scanning, subjects rated the attractiveness of the 40 faces on a scale from 1 to 10. There were no significant differences between ratings by male and female subjects, or between ratings on male and female faces.

Oxyphotobacteria

Antenna ring around photosystem I

The oceanic picoplankton *Prochlorococcus* — probably the most abundant photosynthetic organism on our planet^{1,2} — can grow at great depths where light intensity is very low³. We have found that the chlorophyll-binding proteins in a deep-living strain of this oxyphotobacterium form a ring around a trimer of the photosystem I (PS I) photosynthetic reaction centre, a clever arrangement that maximizes the capture of light energy in such dim conditions.

Many types of oxyphotobacterium collect light for photosynthesis by means of

water-soluble phycobilisomes, but *Prochlorococcus*, *Prochloron* and *Prochlorothrix* all use intrinsic light-collecting proteins that bind to chlorophyll *a* or *b*. Consequently, it was thought that the chloroplasts of green algae and higher plants, which also contain chlorophyll *a/b*-binding proteins, were derived by endosymbiosis from an ancestral prokaryote phylogenetically related to one of these three atypical oxyphotobacteria⁴.

Gene sequencing has shown, however, that the *pcb* genes that encode the principal light-collecting proteins are not closely related to the *cab* genes that encode the chlorophyll *a/b*-binding proteins of green plastids⁵, but instead are similar to the iron-stress-induced *isiA* gene of cyanobacteria⁵. Both encode proteins predicted to have six transmembrane helices that share homology with the six transmembrane helices of CP43, a protein that binds to chlorophyll *a* in photosystem II (PS II).

We investigated how the chlorophyll *a/b*-binding Pcb proteins structurally interact with the PS I reaction centre in *Prochlorococcus*. Figure 1a shows a projection map derived from electron microscopy and single-particle analysis of a large PS I complex isolated from *Prochlorococcus* strain SS120 by centrifugation through a density gradient of sucrose. The central region of the image has threefold symmetry and is surrounded by 18 separate subunits, as inferred from their densities.

This structure is remarkably similar to that of the PS I complex isolated from the cyanobacterium *Synechocystis* PCC6803 after iron deprivation⁶. In this case, the central region consists of a trimer of the PS I reaction centre surrounded by 18 copies of the *isiA* gene product. Given the sequence similarity between the *isiA* and *pcb* genes, the structure shown in Fig. 1a might also represent a trimer of PS I surrounded by 18 copies of one or more of the 7 chlorophyll *a/b*-binding Pcb proteins found in this *Prochlorococcus* strain⁷.

Chlorophyll fluorescence and protein analysis confirmed that this new complex contained PS I and Pcb proteins, and that efficient energy transfer occurs from the Pcb antenna ring to the core of the reaction centre. Furthermore, optical absorption spectra indicated that chlorophyll *a* and *b* are contained within the antenna ring and also within the PS I trimer. Figure 1b, c shows the three-dimensional structure of this Pcb-PS I supercomplex. This was modelled using the X-ray structures of the PS I trimer and CP43 transmembrane helices^{8,9} of the cyanobacterium *Synechococcus elongatus*.

The *Prochlorococcus* strain used here was originally collected from the Sargasso Sea^{2,10}, where it grows at considerable depths and would benefit from an antenna

system of the type reported here in order to increase the collecting capacity of PS I. The ring of Pcb proteins contributes a further 270 light-collecting chlorophyll molecules to the 300 predicted to be associated with the PS I reaction-centre trimer, assuming that the Pcb subunits each bind to 15 chlorophyll molecules.

The antenna ring therefore represents an increase of 90% in light-collecting capacity by the Pcb-PS I supercomplex over that of PS I in cyanobacteria such as *Synechocystis*, when growing in the absence of iron stress. In this latter case, the cyanobacterial PS I/PS II ratio is about 3:1, compared with 1:1 in *Prochlorococcus*. Under iron-stress conditions, however, when the *IsiA* antenna ring forms around PS I, the PS I/PS II ratio drops to about 1:1 in cyanobacteria as well. The formation of these Pcb and *IsiA* rings around PS I may therefore represent a mechanism for increasing the efficiency of this photosystem when its level is close to that of PS II.

It is not clear whether the PS I antenna ring is present in other types of chlorophyll *b*-containing oxyphotobacteria, as the number of *pcb* genes differs in *Prochloron*, *Prochlorothrix* and *Prochlorococcus*, as well as among different strains of *Prochlorococcus*^{5,7}. Although the relationship between gene number and structural features in other strains of *Prochlorococcus* remains to be investigated, our discovery of an antenna ring around PS I in both *Prochlorococcus* and iron-deficient cyanobacteria⁶ opens up discussion on the regulation of light-collection systems in oxyphotobacteria, as well as having implications for the evolution of these photosynthetic prokaryotes. The structures give functional significance to the trimerization of PS I in oxyphotobacteria, as well as providing new systems for studying energy transfer in light-collecting antennae during photosynthesis.

T. S. Bibby*, J. Nield*, F. Partensky†, J. Barber*

*Wolfson Laboratories, Department of Biological Sciences, Imperial College of Science, Technology and Medicine, London SW7 2AY, UK
e-mail: j.barber@ic.ac.uk

†Observatoire Océanologique de Roscoff, Centre National de la Recherche Scientifique et Université Paris 6, Station Biologique, BP74, F-29682 Roscoff Cedex, France

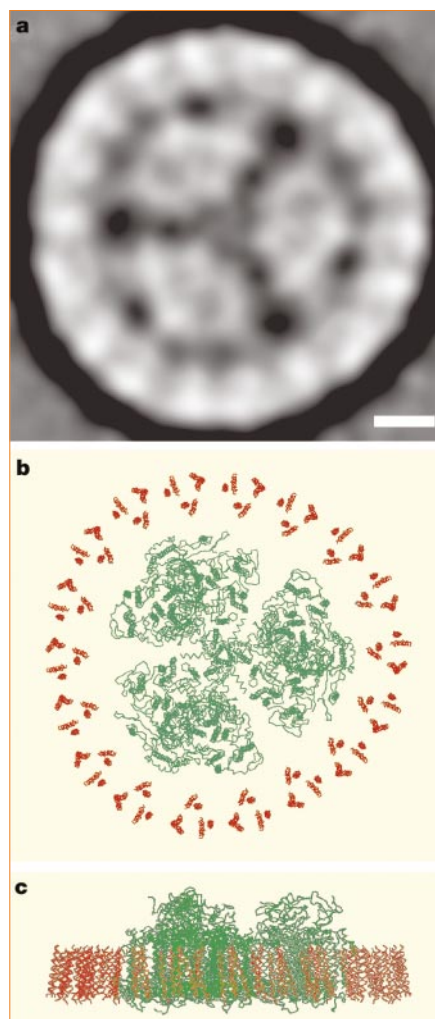


Figure 1 The Pcb-photosystem I (Pcb-PS I) supercomplex of *Prochlorococcus*. **a**, Projection of the top view of the Pcb-PS I supercomplex isolated from *Prochlorococcus marinus*, strain SS120. This map is derived from the initial analyses of 3,500 negatively stained particles viewed under electron microscopy at room temperature (scale bar, 5 nm). **b**, **c**, Top view from the stromal side (**b**), and side view showing the extrinsic domains on the stromal side (**c**) of a three-dimensional model of the Pcb-PS I supercomplex; the model is based on X-ray data from cyanobacterial PS I trimers¹⁰ (green) and CP43 transmembrane helices of cyanobacterial photosystem II (ref. 8; red).

- Chisholm, S. W. et al. *Nature* **334**, 340–343 (1988).
- Chisholm, S. W. et al. *Arch. Microbiol.* **157**, 297–300 (1992).
- Partensky, F., Hess, W. R. & Vaulot, D. *Microbiol. Mol. Biol. Rev.* **63**, 106–127 (1999).
- Lewin, R. A. *Nature* **261**, 697–698 (1976).
- La Roche, J. et al. *Proc. Natl Acad. Sci. USA* **93**, 15244–15248 (1996).
- Bibby, T. S., Nield, J. & Barber, J. *Nature* **412**, 743–745 (2001).
- Garczarek, L. et al. *Proc. Natl Acad. Sci. USA* **97**, 4098–4101 (2000).
- Zouni, A. et al. *Nature* **409**, 739–743 (2001).
- Krauss, N. et al. *Nature Struct. Biol.* **3**, 965–973 (1996).
- Moore, L. R., Goericke, R. & Chisholm, S. W. *Ecol. Prog. Ser.* **116**, 259–275 (1995).

Catastrophic shifts in ecosystems

Marten Scheffer*, Steve Carpenter†, Jonathan A. Foley‡, Carl Folke§ & Brian Walker||

* Department of Aquatic Ecology and Water Quality Management, Wageningen University, PO Box 8080, NL-6700 DD Wageningen, The Netherlands

† Center for Limnology, University of Wisconsin, 680 North Park Street, Madison, Wisconsin 53706, USA

‡ Center for Sustainability and the Global Environment (SAGE), Institute for Environmental Studies, University of Wisconsin, 1225 West Dayton Street, Madison, Wisconsin 53706, USA

§ Department of Systems Ecology and Centre for Research on Natural Resources and the Environment (CNM), Stockholm University, S-10691 Stockholm, Sweden

|| CSIRO Sustainable Ecosystems, GPO Box 284, Canberra, Australian Capital Territory 2601, Australia

All ecosystems are exposed to gradual changes in climate, nutrient loading, habitat fragmentation or biotic exploitation. Nature is usually assumed to respond to gradual change in a smooth way. However, studies on lakes, coral reefs, oceans, forests and arid lands have shown that smooth change can be interrupted by sudden drastic switches to a contrasting state. Although diverse events can trigger such shifts, recent studies show that a loss of resilience usually paves the way for a switch to an alternative state. This suggests that strategies for sustainable management of such ecosystems should focus on maintaining resilience.

The notion that ecosystems may switch abruptly to a contrasting alternative stable state emerged from work on theoretical models^{1,2}. Although this provided an inspiring search image for ecologists, the first experimental examples that were proposed were criticized strongly³. Indeed, it seemed easier to demonstrate shifts between alternative stable states in models than in the real world. In particular, unravelling the mechanisms governing the behaviour of spatially extensive ecosystems is notoriously difficult, because it requires the interfacing of phenomena that occur on very different scales of space, time and ecological organization⁴. Nonetheless, recent studies have provided a strong case for the existence of alternative stability domains in various important ecosystems^{5–8}. Here, we do not address brief switches to alternative states such as described for pest outbreaks⁹. Also, we do not fully cover the extensive work on positive feedbacks and multiple stable states in ecological systems. Instead, we concentrate on observed large-scale shifts in major ecosystems and their explanations. After sketching the theoretical framework, we present an overview of results from different ecosystems, highlight emerging patterns, and discuss how these insights may contribute to improved management.

Theoretical framework

Ecosystem response to gradually changing conditions

External conditions to ecosystems such as climate, inputs of nutrients or toxic chemicals, groundwater reduction, habitat fragmentation, harvest or loss of species diversity often change gradually, even linearly, with time^{10,11}. The state of some ecosystems may respond in a smooth, continuous way to such trends (Fig. 1a). Others may be quite inert over certain ranges of conditions, responding more strongly when conditions approach a certain critical level (Fig. 1b). A crucially different situation arises when the ecosystem response curve is 'folded' backwards (Fig. 1c). This implies that, for certain environmental conditions, the ecosystem has two alternative stable states, separated by an unstable equilibrium that marks the border between the 'basins of attraction' of the states.

The presence of alternative stable states has profound implications for the response to environmental change (Fig. 2a). When the ecosystem is in a state on the upper branch of the folded curve, it can not pass to the lower branch smoothly. Instead, when conditions change sufficiently to pass the threshold ('saddle-node' or 'fold' bifurcation, F_2), a 'catastrophic' transition to the lower branch occurs. Note that when one monitors the system on a stable branch before a switch, little change in its state is observed. Indeed, such catastrophic shifts

occur typically quite unannounced, and 'early-warning signals' of approaching catastrophic change are difficult to obtain. Another important feature is that to induce a switch back to the upper branch, it is not sufficient to restore the environmental conditions of before the collapse (F_2). Instead, one needs to go back further, beyond the other switch point (F_1), where the system recovers by shifting back to the upper branch. This pattern, in which the forward and backward switches occur at different critical conditions, is known as hysteresis. The degree of hysteresis may vary strongly even in the same kind of ecosystem. For instance, shallow lakes can have a pronounced hysteresis in response to nutrient loading (Fig. 1c), whereas deeper lakes may react smoothly (Fig. 1b)¹². A range of mathematical models of specific ecological systems with alternative stable states has been published. Box 1 shows an example of a simple model that can be thought of as describing desertification or lake eutrophication.

Effects of stochastic events

In the real world, conditions are never constant. Stochastic events such as weather extremes, fires or pest outbreaks can cause fluctuations in the conditioning factors (horizontal axis) but often affect the state (vertical axis) directly, for example, by wiping out parts of populations. If there is only one basin of attraction, the system will settle back to essentially the same state after such events. However, if there are alternative stable states, a sufficiently severe perturbation of the ecosystem state may bring the system into the basin of attraction of another state (Fig. 2b). The likelihood of this depends not only on the perturbation, but also on the size of the attraction basin. In terms of stability landscapes (Fig. 3), if the valley is small, a small perturbation may be enough to displace the ball far enough to push it over the hill, resulting in a shift to the alternative stable state. Following Holling¹, we here use the term 'resilience' to refer the size of the valley, or basin of attraction, around a state, which corresponds to the maximum perturbation that can be taken without causing a shift to an alternative stable state.

In systems with multiple stable states, gradually changing conditions may have little effect on the state of the ecosystem, but nevertheless reduce the size of the attraction basin (Fig. 3). This loss of resilience makes the system more fragile in the sense that can easily be tipped into a contrasting state by stochastic events. Such stochastic fluctuations may often be driven externally; however, they can also result from internal system dynamics. The latter can happen if the alternative

attractors are 'cycles' or 'strange attractors', rather than equilibria. A system that moves along a strange attractor fluctuates chaotically even in the absence of an external stochastic forcing. These fluctuations can lead to a collision with the boundary of the basin of attraction, and consequently induce a switch to an alternative state. Models indicate that such 'non-local bifurcations'¹³ or 'basin boundary collisions'¹⁴ may occur in ocean-climate systems¹⁵ as well as various ecosystems⁹. In practice, it will often be a blend of internal processes and external forcing that generates fluctuations¹⁶ that can induce a state shift by bringing systems with reduced resilience over the boundary of an attraction basin. In view of these permanent fluctuations, the term 'stable state' is hardly appropriate for any ecosystem. Nonetheless, for the sake of clarity we use 'state' rather than the more correct term 'dynamic regime'.

Examples

Lakes

Shifts between alternative stable states occur in lakes^{12,17}. One of the best-studied and most dramatic state shifts is the sudden

loss of transparency and vegetation observed in shallow lakes subject to human-induced eutrophication^{5,18}. The pristine state of most shallow lakes is probably one of clear water and a rich submerged vegetation. Nutrient loading has changed this situation in many cases. Remarkably, water clarity often seems to be hardly affected by increased nutrient concentrations until a critical threshold is passed, at which the lake shifts abruptly from clear to turbid. With this increase in turbidity, submerged plants largely disappear. Associated loss of animal diversity and reduction of the high algal biomass makes this state undesired. Reduction of nutrient concentrations is often insufficient to restore the vegetated clear state. Indeed, the restoration of clear water happens at substantially lower nutrient levels than those at which the collapse of the vegetation occurred (Fig. 4). Experimental work suggests that plants increase water clarity, thereby enhancing their own growing conditions⁵. This causes the clear state to be a self-stabilizing alternative to the turbid situation (Fig. 5). The reduction of phytoplankton biomass and turbidity by vegetation involves a suite of mechanisms, including reduction of nutrients in the water column, protection of phytoplankton grazers such as *Daphnia* against fish predation, and prevention of sediment resuspension. In contrast, fish are central in maintaining the turbid state, because they control *Daphnia* in the absence of plants. Also, in search for benthic food they resuspend sediments, adding to turbidity. Whole-lake experiments show that a temporary strong reduction of fish biomass as 'shock therapy' can bring such lakes back into a permanent clear state if the nutrient level is not too high¹⁹.

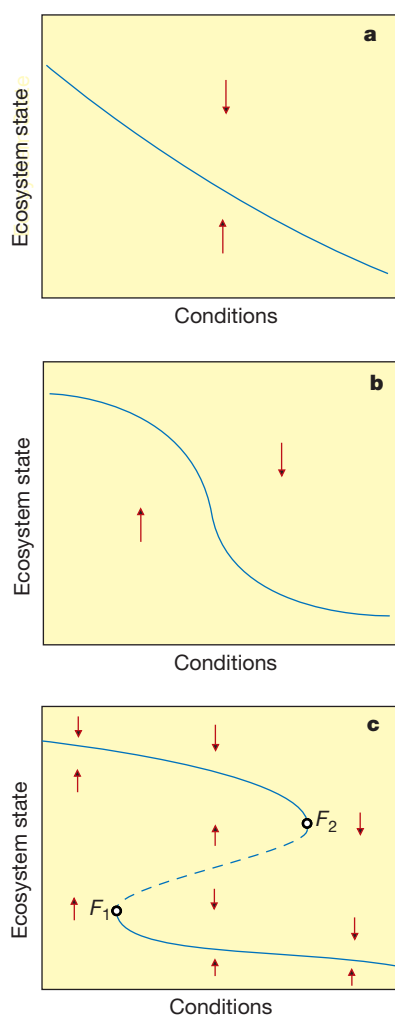


Figure 1 Possible ways in which ecosystem equilibrium states can vary with conditions such as nutrient loading, exploitation or temperature rise. In **a** and **b**, only one equilibrium exists for each condition. However, if the equilibrium curve is folded backwards (**c**), three equilibria can exist for a given condition. It can be seen from the arrows indicating the direction of change that in this case equilibria on the dashed middle section are unstable and represent the border between the basins of attraction of the two alternative stable states on the upper and lower branches. Modified from ref. 58.

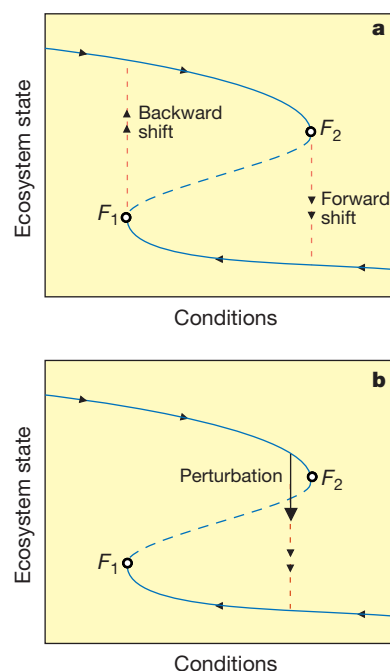


Figure 2 Two ways to shift between alternative stable states. **a**, If the system is on the upper branch, but close to the bifurcation point F_2 , a slight incremental change in conditions may bring it beyond the bifurcation and induce a catastrophic shift to the lower alternative stable state ('forward shift'). If one tries to restore the state on the upper branch by means of reversing the conditions, the system shows hysteresis. A backward shift occurs only if conditions are reversed far enough to reach the other bifurcation point, F_1 . **b**, A perturbation (arrow) may also induce a shift to the alternative stable state, if it is sufficiently large to bring the system over the border of the attraction basin (see also Fig. 3).

Coral reefs

Coral reefs are known for their high biodiversity. However, many reefs around the world have degraded. A major problem is that corals are overgrown by fleshy macroalgae. Reef ecosystems seem to shift between alternative stable states, rather than responding in a smooth way to changing conditions^{20–22}. The shift to algae in Caribbean reefs is the result of a combination of factors that make the system vulnerable to events that trigger the actual shift⁸. These factors presumably include increased nutrient loading as a result of changed land-use and intensive fishing, which reduced the numbers of large fish and subsequently of the smaller herbivorous species. Consequently, the sea urchin *Diadema antillarum*, which competes with the herbivorous fish for algal food, increased in numbers. In 1981 a hurricane caused extensive damage to the reefs, but despite high nutrient levels, algae invading the open areas were controlled by *Diadema*, allowing coral to recolonize. However, in subsequent years, populations of *Diadema* were dramatically reduced by a pathogen. Because herbivorous fish had also become rare, algae were released from the control of grazers and the reefs rapidly became overgrown by fleshy brown algae. This switch is thought to be difficult to reverse because adult algae are less palatable to herbivores and the algae prevent settlement of coral larvae.

Woodlands

Many studies indicate that woodlands and a grassy open landscape can be alternative stable states. Landscapes can be kept open by herbivores (often in combination with fires) because seedlings of woody plants, unlike adult trees, are easily eliminated by herbivores. Conversely, woodlands, once established, are stable because adult trees can not be destroyed by herbivores and shading reduces grass cover so that fires can not spread. Well analysed examples are African woodland dynamics in Botswana²³ and Tanzania²⁴, where regeneration of woodlands occurred for a few decades from the 1890s because of low herbivore numbers due to a combination of rinderpest epidemic and elephant hunting. Once established, these woodlands could not be eliminated by grazers. However, the current destruction of woodlands by humans and high densities of elephants is probably irreversible in these regions

Box 1

A minimal mathematical model

A minimal model of an ecosystem showing hysteresis describes the change over time of an 'unwanted' ecosystem property x :

$$dx/dt = a - bx + rf(x) \quad (1)$$

The parameter a represents an environmental factor that promotes x . The remainder of the equation describes the internal dynamics: b represents the rate at which x decays in the system, whereas r is the rate at which x recovers again as a function f of x . For lakes, one can think of x as nutrients suspended in phytoplankton causing turbidity, of a as nutrient loading, of b as nutrient removal rate and of r as internal nutrient recycling¹². For desertification, one could interpret x as barren soil, a as vegetation destruction, b as recolonization of barren soil by plants and r as erosion by wind and runoff⁶⁸.

For $r = 0$, the model has a single equilibrium at $x = a/b$. The last term, however, can cause alternative stable states, for example, if $f(x)$ is a function that increases steeply at a threshold h , as in the case of the Hill function:

$$f(x) = x^p / (x^p + h^p)$$

where the exponent p determines the steepness of the switch occurring around h . Notice that (1) can have multiple stable states only if the maximum $\{rf'(x)\} > b$. Thus, steeper Hill functions (resulting from higher p values) create stronger hysteresis.

unless herbivore numbers are again reduced (which is unlikely given the focus of the national parks' policy on attracting tourists)²³.

In dry areas, conditions in the absence of cover by adult trees may be too desiccating to allow the seedlings to survive, even in the absence of herbivores, implying a more severe irreversibility of woodland loss, as illustrated by matorral woodlands in the drier parts of Mediterranean central Chile²⁵. This implies that only rare combinations of wet years and repressed herbivore populations may allow recovery of these diverse woodlands, which once covered extensive areas. Another case of irreversible loss of trees is that of cloud forests²⁶. Condensation of water from clouds in the canopy supplies moisture for a rich ecosystem. If the trees are cut, this water input stops and the resulting conditions can be too dry for recovery of the forest.

In savannahs, sparse trees with a grass layer are the natural state. A shift to a dense woody state (known as 'bush encroachment') can result from a combination of change in fire and grazing regimes. Occasional natural fires reduce the woody plant cover and favour development of the grass layer. However, excessive grazing by livestock reduces grass and hence fuel for fire. In the absence of fire, cohorts of shrubs establish during wet years and can suppress grass cover, thereby inhibiting the spread of fire. The system stays in this thicket state until trees begin to die, thereby allowing grass cover to attain levels that will carry an effective fire^{27,28}.

Deserts

Desertification—the loss of perennial vegetation in arid and semi-arid regions—is often cited as one of the main ecological threats facing the world today²⁹, although the pace at which it proceeds in the Sahara region may be less than previously

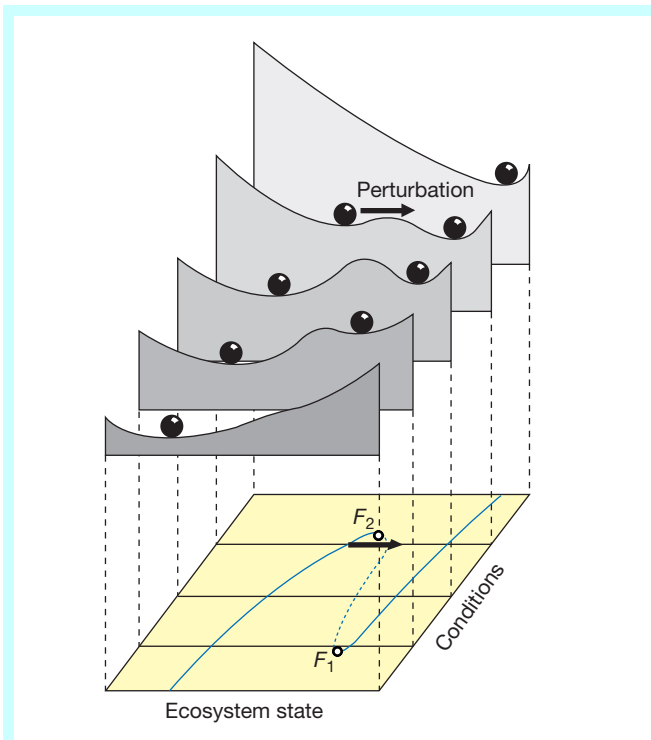


Figure 3 External conditions affect the resilience of multi-stable ecosystems to perturbation. The bottom plane shows the equilibrium curve as in Fig. 2. The stability landscapes depict the equilibria and their basins of attraction at five different conditions. Stable equilibria correspond to valleys; the unstable middle section of the folded equilibrium curve corresponds to a hill. If the size of the attraction basin is small, resilience is small and even a moderate perturbation may bring the system into the alternative basin of attraction.

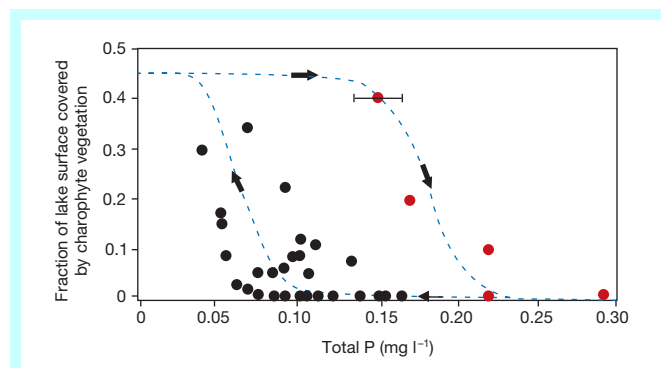


Figure 4 Hysteresis in the response of charophyte vegetation in the shallow Lake Veluwe to increase and subsequent decrease of the phosphorus concentration. Red dots represent years of the forward switch in the late 1960s and early 1970s. Black dots show the effect of gradual reduction of the nutrient loading leading eventually to the backward switch in the 1990s. From ref. 59.

thought³⁰. Various lines of evidence indicate that vegetated and desert situations may represent alternative stable states. Local soil–plant interactions are important in determining the stability of perennial plant cover^{6,31}. Perennial vegetation allows precipitation to be absorbed by the topsoil and to become available for uptake by plants. When vegetation cover is lost, runoff increases, and water entering the soil quickly disappears to deeper layers where it cannot be reached by most plants. Wind and runoff also erode fertile remains of the topsoil, making the desert state even more hostile for recolonizing seedlings. As a result, the desert state can be too harsh to be recolonized by perennial plants, even though a perennial vegetation may persist once it is present, owing to the enhancement of soil conditions.

On a much larger scale, a feedback between vegetation and climate may also lead to alternative stable states. The Sahel region seems to shift back and forth between a stable dry and a stable moister climatic regime. For example, every year since 1970 has been anomalously dry, whereas every year of the 1950s was unusually wet; in other parts of the world, runs of wet or

dry years typically do not exceed 2–5 years³². Many studies have addressed the question of why this system shifts between distinct modes, instead of drifting through a series of intermediate conditions. A new generation of coupled climate–ecosystem models^{33–35} demonstrates that Sahel vegetation itself may have a role in the drought dynamics, especially in maintaining long periods of wet or dry conditions. The mechanism is one of positive feedback: vegetation promotes precipitation and vice versa, leading to alternative states.

Intriguing evidence for alternative stable states in the Sahel and Sahara desert systems comes from ancient abrupt shifts at a large scale between desert and vegetated states, coupled to climatic change in North Africa. During the early and middle Holocene—about 10,000 to 5,000 years before present—much of the Sahara was wetter than it is today, with extensive vegetation cover and lakes and wetlands^{36,37}. Then, some time around 5,000 years before present, an abrupt switch to desert-like conditions occurred³⁸. By means of combined atmosphere–ocean–biosphere models, it has been shown that feedbacks causing alternative stable states could indeed explain such an abrupt switch, even when the climate system is being driven by slow gradual change in insolation resulting from subtle variations in the Earth's orbit (Fig. 6)^{38,39}.

The timescales in this example are rather long. Nonetheless, it illustrates the same phenomenon of alternative stability domains that underlies the dynamics found in the other examples. An important implication here is that small environmental changes, such as overgrazing⁴⁰, increased dust loading³², or changes in nearby ocean temperatures³³, may potentially cause a total state shift for the entire area once a certain critical threshold is passed.

Oceans

Time series of fish catches, oyster condition, plankton abundance and other marine ecosystem properties indicate conspicuous jumps from one rather stable condition to another (Fig. 7). These puzzling events have been termed 'regime shifts'⁴¹. The implications of oceanic regime shifts for fisheries and oceanic CO₂ uptake⁴² are profound, but the cause of the shifts is poorly understood⁴¹. In view of the overriding importance of sea currents on these ecosystems, changes in the oceanic circulation or weather pattern can reasonably be

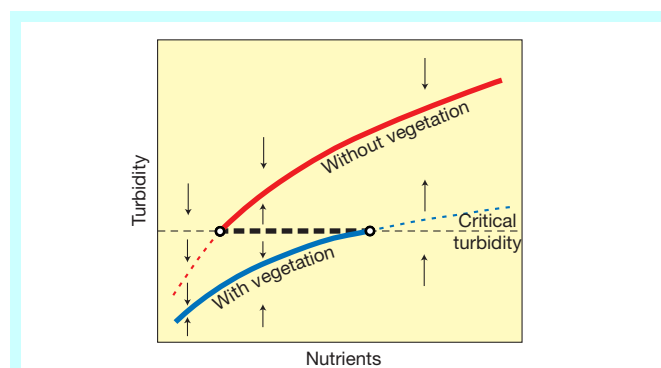


Figure 5 A graphical model⁶⁰ of alternative stable states in shallow lakes on the basis of three assumptions: (1) turbidity of the water increases with the nutrient level; (2) submerged vegetation reduces turbidity; and (3) vegetation disappears when a critical turbidity is exceeded. In view of the first two assumptions, equilibrium turbidity can be drawn as two different functions of the nutrient level: one for a vegetation-dominated situation, and one for an unvegetated situation. Above a critical turbidity, vegetation will be absent, in which case the upper equilibrium line is the relevant one; below this turbidity the lower equilibrium curve applies. As a result, at lower nutrient levels, only the vegetation-dominated equilibrium exists, whereas at the highest nutrient levels, there is only an unvegetated equilibrium. Over a range of intermediate nutrient levels, two alternative equilibria exist: one with vegetation, and a more turbid one without vegetation, separated by a (dashed) unstable equilibrium.

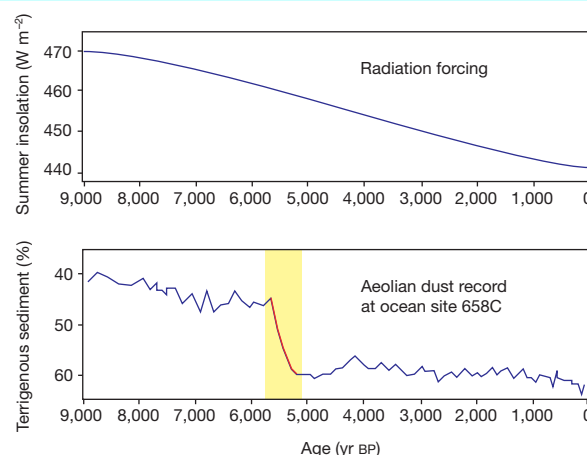


Figure 6 Over the past 9,000 years, average Northern Hemisphere summer insolation (upper panel) has varied gradually owing to subtle variation in the Earth's orbit. About 5,000 years before present (yr BP), this change in solar radiation triggered an abrupt shift in climate and vegetation cover over the Sahara, as reflected in the contribution of terrigenous (land-eroded) dust to oceanic sediment at a sample site near the African coast (lower panel). Modified from ref. 61.

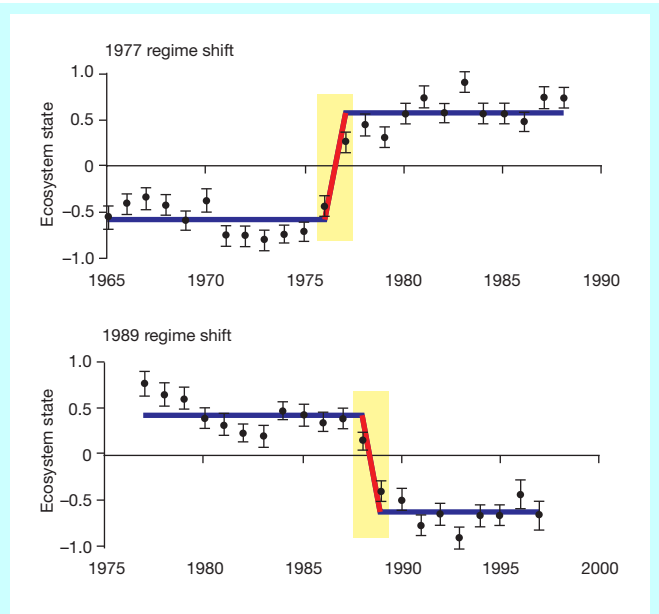


Figure 7 Distinct state shifts occurred in the Pacific Ocean ecosystem around 1977 and 1989. The compound indices of ecosystem state are obtained by averaging 31 climatic and 69 biological normalized time series. Modified from ref. 41.

expected to be the drivers of change. However, the state shifts are sometimes reflected more consistently by the biological data than by the physical indices, suggesting that biotic feedbacks could be stabilizing the community in a certain state, and that shifts to a different state are triggered merely by physical events⁴¹. It is becoming increasingly clear that competition and predation are much more important in driving oceanic community dynamics than previously thought⁴³. It is therefore not surprising that fisheries can affect the entire food web, causing profound shifts in species abundance on various trophic levels^{44–46}. Also, such tight biotic interactions imply that sensitivity of a single keystone species to subtle environmental change can cause major shifts in community composition⁴⁷. Therefore, solving the puzzle of regime shifts in oceanic ecosystems may require unravelling the interplay of effects of fisheries and effects of changes in the physical climate or ocean system.

The coupled ocean–climate system may also go through shifts between alternative stable states that are much more drastic than the regime shifts mentioned above^{15,48}. For example, simulation studies indicate that gradual climate warming may cause an increase in freshwater inflow into the North Atlantic that

prevents the formation of dense deep water, which is needed to power the ‘global conveyor belt’ oceanic current that transports warm water to eastern North America and western Europe¹⁵. Such a change causes the climate in these regions to become dramatically colder. Reconstructions of palaeoclimate show that similar large shifts have happened in the past and can be very swift indeed, occurring in less than a decade⁴⁸.

Emerging patterns

All of these case studies suggest shifts between alternative stable states. Nonetheless, proof of multiplicity of stable states is usually far from trivial. Observation of a large shift *per se* is not sufficient, as systems may also respond in a nonlinear way to gradual change if they have no alternative stable states (for example, as in Fig. 1b)⁴⁹. Also, the power of statistical methods to infer the underlying system properties from noisy time series is poor^{7,50,51}. However, mere demonstration of a positive-feedback mechanism is also insufficient as proof of alternative stable states, because it leaves a range of possibilities between pronounced hysteresis and smooth response, depending on the strength of the feedback and other factors⁴⁹. Indeed, the strongest cases for the existence of alternative stable states are based on combinations of approaches, such as observations of repeated shifts, studies of feedback mechanisms that tend to maintain the different states, and models showing that these mechanisms can plausibly explain field data.

Although the specific details of the reviewed state shifts differ widely, an overview (Table 1) shows some consistent patterns. First, the contrast among states in ecosystems is usually due to a shift in dominance among organisms with different life forms. Second, state shifts are usually triggered by obvious stochastic events such as pathogen outbreaks, fires or climatic extremes. Third, feedbacks that stabilize different states involve both biological and physical and chemical mechanisms.

Perhaps most importantly, all models of ecosystems with alternative stable states indicate that gradual change in environmental conditions, such as human-induced eutrophication and global warming, may have little apparent effect on the state of these systems, but still alter the ‘stability domain’ or resilience of the current state and hence the likelihood that a shift to an alternative state will occur in response to natural or human-induced fluctuations.

Implications for management

Ecosystem state shifts can cause large losses of ecological and economic resources, and restoring a desired state may require drastic and expensive intervention⁵². Thus, neglect of the possibility of shifts to alternative stable states in ecosystems may have heavy costs to society. Because of hysteresis in their

Table 1 Characteristics of some major ecosystem state shifts and their causes

Ecosystem	State I	State II	Events inducing shift from I to II	Events inducing shift from II to I	Suggested main causes of hysteresis	Factors affecting resilience
Lakes	Clear with submerged vegetation	Turbid with phytoplankton	Killing of plants by herbicide Killing of <i>Daphnia</i> by pesticide High water level	Killing of fish Low water level	Positive feedback of plant growth Trophic feedbacks	Nutrient accumulation
Coral reefs	Corals	Fleshy brown macroalgae	Killing of coral by hurricane Killing of sea urchins by pathogen	Unknown	Prevention of coral recolonization by unpalatable adult algae	Nutrient accumulation Climate change Fishing
Woodlands	Herbaceous vegetation	Woodlands	Fires Tree cutting	Killing of grazers by pathogen Hunting of grazers	Positive feedback of plant growth Inedibility of adult trees	Overgrazing Climate change
Deserts	Perennial vegetation	Bare soil with ephemeral plants	Climatic events Overgrazing by cattle	Climatic events	Positive feedback of plant growth	Climate change
Oceans	Various	Various	Climatic events	Climatic events	Physical	Fishing Climate change

response and the invisibility of resilience itself, these systems typically lack early-warning signals of massive change. Therefore attention tends to focus on precipitating events rather than on the underlying loss of resilience. For example, gradual changes in the agricultural watershed increased the vulnerability of Lake Apopka (Florida, USA) to eutrophication, but a hurricane wiped out aquatic plants in 1947 and probably triggered the collapse of water quality^{53,54}; gradual increase in nutrient inputs and fishing pressure created the potential for algae to overgrow Caribbean corals, but overgrowth was triggered by a conspicuous disease outbreak among sea urchins that released algae from grazer control⁸; and gradual increase in grazing decreases the capacity of Australian rangelands to carry the fires that normally control shrubs, but extreme wet years trigger the actual shift to shrub dominance^{27,55}.

Prevention of perturbations is often a major goal of ecosystem management, not surprisingly. This is unfortunate, not only because disturbance is a natural component of ecosystems that promotes diversity and renewal processes^{56,57}, but also because it distracts attention from the underlying structural problem of resilience. The main implication of the insights presented here is that efforts to reduce the risk of unwanted state shifts should address the gradual changes that affect resilience rather than merely control disturbance. The challenge is to sustain a large stability domain rather than to control fluctuations. Stability domains typically depend on slowly changing variables such as land use, nutrient stocks, soil properties and biomass of long-lived organisms. These factors may be predicted, monitored and modified. In contrast, stochastic events that trigger state shifts (such as hurricanes, droughts or disease outbreaks) are usually difficult to predict or control. Therefore, building and maintaining resilience of desired ecosystem states is likely to be the most pragmatic and effective way to manage ecosystems in the face of increasing environmental change. □

- Holling, C. S. Resilience and stability of ecological systems. *Annu. Rev. Ecol. Syst.* **4**, 1–23 (1973).
- May, R. M. Thresholds and breakpoints in ecosystems with a multiplicity of stable states. *Nature* **269**, 471–477 (1977).
- Connell, J. H. & Sousa, W. P. On the evidence needed to judge ecological stability or persistence. *Am. Nat.* **121**, 789–824 (1983).
- Levin, S. A. The problem of pattern and scale in ecology. *Ecology* **73**, 1943–1967 (1992).
- Scheffer, M., Hosper, S. H., Meijer, M. L. & Moss, B. Alternative equilibria in shallow lakes. *Trends Ecol. Evol.* **8**, 275–279 (1993).
- Van de Koppel, J., Rietkerk, M. & Weissing, F. J. Catastrophic vegetation shifts and soil degradation in terrestrial grazing systems. *Trends Ecol. Evol.* **12**, 352–356 (1997).
- Carpenter, S. R. in *Ecology: Achievement and Challenge* (eds Press, M. C., Huntly, N. & Levin, S.) (Blackwell, London, 2001).
- Nystrom, M., Folke, C. & Moberg, F. Coral reef disturbance and resilience in a human-dominated environment. *Trends Ecol. Evol.* **15**, 413–417 (2000).
- Rinaldi, S. & Scheffer, M. Geometric analysis of ecological models with slow and fast processes. *Ecosystems* **3**, 507–521 (2000).
- Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. Human domination of Earth's ecosystems. *Science* **277**, 494–499 (1997).
- Tilman, D. et al. Forecasting agriculturally driven global environmental change. *Science* **292**, 281–284 (2001).
- Carpenter, S. R., Ludwig, D. & Brock, W. A. Management of eutrophication for lakes subject to potentially irreversible change. *Ecol. Appl.* **9**, 751–771 (1999).
- Kuznetsov, Y. A. *Elements of Applied Bifurcation Theory* (Springer, New York, 1995).
- Vandermeer, J. & Yodzis, P. Basin boundary collision as a model of discontinuous change in ecosystems. *Ecology* **80**, 1817–1827 (1999).
- Rahmstorf, S. Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle. *Nature* **378**, 145–149 (1995); erratum **379**, 847 (1996).
- Ellner, S. & Turchin, P. Chaos in a noisy world: New methods and evidence from time-series analysis. *Am. Nat.* **145**, 343–375 (1995).
- Scheffer, M., Rinaldi, S., Gragnani, A., Mur, L. R. & Van Nes, E. H. On the dominance of filamentous cyanobacteria in shallow, turbid lakes. *Ecology* **78**, 272–282 (1997).
- Jeppesen, E. et al. Lake and catchment management in Denmark. *Hydrobiologia* **396**, 419–432 (1999).
- Meijer, M. L., Jeppesen, E., Van Donk, E. & Moss, B. Long-term responses to fish-stock reduction in small shallow lakes: Interpretation of five-year results of four biomanipulation cases in the Netherlands and Denmark. *Hydrobiologia* **276**, 457–466 (1994).
- Knowlton, N. Thresholds and multiple stable states in coral reef community dynamics. *Am. Zool.* **32**, 674–682 (1992).

- Done, T. J. Phase shifts in coral reef communities and their ecological significance. *Hydrobiologia* **247**, 121–132 (1991).
- McCook, L. J. Macroalgae, nutrients and phase shifts on coral reefs: Scientific issues and management consequences for the Great Barrier Reef. *Coral Reefs* **18**, 357–367 (1999).
- Walker, B. H. in *Conservation Biology for the Twenty-first Century* (eds Weston, D. & Pearl, M.), 121–130 (Oxford Univ. Press, Oxford, 1989).
- Dublin, H. T., Sinclair, A. R. & McGlade, J. Elephants and fire as causes of multiple stable states in the Serengeti–Mara woodlands. *J. Anim. Ecol.* **59**, 1147–1164 (1990).
- Holmgren, M. & Scheffer, M. El Niño as a window of opportunity for the restoration of degraded arid ecosystems. *Ecosystems* **4**, 151–159 (2001).
- Wilson, J. B. & Agnew, A. D. Q. Positive-feedback switches in plant communities. *Adv. Ecol. Res.* **23**, 263–336 (1992).
- Walker, B. H. Rangeland ecology: understanding and managing change. *Ambio* **22**, 2–3 (1993).
- Ludwig, D., Walker, B. & Holling, C. S. Sustainability, stability and resilience. *Conserv. Ecol.* [online] (01 Aug. 01) (<http://www.consecol.org/vol1/iss1/art7/>) (1997).
- Kassas, M. Desertification: A general review. *J. Arid Environ.* **30**, 115–128 (1995).
- Tucker, C. J. & Nicholson, S. E. Variations in the size of the Sahara Desert from 1980 to 1997. *Ambio* **28**, 587–591 (1999).
- Rietkerk, M., Van den Bosch, F. & Van de Koppel, J. Site-specific properties and irreversible vegetation changes in semi-arid grazing systems. *Oikos* **80**, 241–252 (1997).
- Nicholson, S. E. Land surface processes and Sahel climate. *Rev. Geophys.* **38**, 117–139 (2000).
- Zeng, N., Neelin, J. D., Lau, K. M. & Tucker, C. J. Enhancement of interdecadal climate variability in the Sahel by vegetation interaction. *Science* **286**, 1537–1540 (1999).
- Wang, G. L. & Eltahir, E. B. Ecosystem dynamics and the Sahel drought. *Geophys. Res. Lett.* **27**, 795–798 (2000).
- Wang, G. L. & Eltahir, E. B. Role of vegetation dynamics in enhancing the low-frequency variability of the Sahel rainfall. *Water Resour. Res.* **36**, 1013–1021 (2000).
- Hoelzmann, P. et al. Mid-Holocene land-surface conditions in northern Africa and the Arabian Peninsula: A data set for the analysis of biogeophysical feedbacks in the climate system. *Global Biogeochem. Cy.* **12**, 35–51 (1998).
- Jolly, D. et al. Biome reconstruction from pollen and plant macrofossil data for Africa and the Arabian peninsula at 0 and 6000 years. *J. Biogeog.* **25**, 1007–1027 (1998).
- Clausen, M. et al. Simulation of an abrupt change in Saharan vegetation in the mid-Holocene. *Geophys. Res. Lett.* **26**, 2037–2040 (1999).
- Brovin, V., Clausen, M., Petoukhov, V. & Ganopolski, A. On the stability of the atmosphere–vegetation system in the Sahara/Sahel region. *J. Geophys. Res.—Atmos.* **103**, 31613–31624 (1998).
- Charney, J. G. The dynamics of deserts and droughts. *J. R. Meteorol. Soc.* **101**, 193–202 (1975).
- Hare, S. R. & Mantua, N. J. Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Prog. Oceanogr.* **47**, 103–145 (2000).
- Reid, P. C., Edwards, M., Hunt, H. G. & Warner, A. J. Phytoplankton change in the North Atlantic. *Nature* **391**, 546–546 (1998).
- Verity, P. G. & Smetacek, V. Organism life cycles, predation, and the structure of marine pelagic ecosystems. *Mar. Ecol. Prog. Ser.* **130**, 277–293 (1996).
- Cury, P. et al. Small pelagics in upwelling systems: Patterns of interaction and structural changes in “wasp-waist” ecosystems. *ICES J. Mar. Sci.* **57**, 603–618 (2000).
- Shiomoto, A., Tadokoro, K., Nagasawa, K. & Ishida, Y. Trophic relations in the subarctic North Pacific ecosystem: Possible feeding effect from pink salmon. *Mar. Ecol. Prog. Ser.* **150**, 75–85 (1997).
- Reid, P. C., Battle, E. -J. V., Batten, S. D. & Brander, K. M. Impacts of fisheries on plankton community structure. *ICES J. Mar. Sci.* **57**, 495–502 (2000).
- Hall, C. A. S., Stanford, J. A. & Hauer, F. R. The distribution and abundance of organisms as a consequence of energy balance along multiple environmental gradients. *Oikos* **65**, 377–390 (2000).
- Taylor, K. Rapid climate change. *Am. Sci.* **87**, 320–327 (1999).
- Scheffer, M. *Ecology of Shallow Lakes* (Chapman and Hall, London, 1998).
- Carpenter, S. R. & Pace, M. L. Dystrophy and eutrophy in lake ecosystems: Implications of fluctuating inputs. *Oikos* **78**, 3–14 (1997).
- Ives, A. R. & Jansen, V. A. A. Complex dynamics in stochastic tritrophic models. *Ecology* **79**, 1039–1052 (1998).
- Maler, K. G. Development, ecological resources and their management: A study of complex dynamic systems. *Eur. Econ. Rev.* **44**, 645–665 (2000).
- Schelske, C. L. in *Proc. 14th Diatom Symp. 1996* (eds Mayama, S., Idei, M. & Koizumi, I.) 367–382 (Koeltz, Koenigstein, 1999).
- Schelske, C. L. & Brezonik, P. in *Restoration of Aquatic Ecosystems* (eds Maurizi, S. & Poillon, F.) 393–398 (National Academic Press, Washington DC, 1992).
- Tongway, D. & Ludwig, J. in *Landscape Ecology, Function and Management: Principles from Australia's Rangelands* (eds Ludwig, J., Tongway, D., Freudenberger, D., Noble, J. & Hodgkinson, K.) 49–61 (CSIRO, Melbourne, 1997).
- Holling, C. S. & Meffe, G. K. Command and control and the pathology of natural resource management. *Cons. Biol.* **10**, 328–337 (1996).
- Paine, R. T., Tegner, M. J. & Johnson, E. A. Compounded perturbations yield ecological surprises. *Ecosystems* **1**, 535–545 (1998).
- Scheffer, M., Brock, W. & Westley, F. Socioeconomic mechanisms preventing optimum use of ecosystem services: an interdisciplinary theoretical analysis. *Ecosystems* **3**, 451–471 (2000).
- Meijer, M. L. *Biomanipulation in the Netherlands—15 Years of Experience*. 1–208 (Wageningen Univ., Wageningen, 2000).
- Scheffer, M. Multiplicity of stable states in freshwater systems. *Hydrobiologia* **200/201**, 475–486 (1990).
- deMenocal, P. et al. Abrupt onset and termination of the African Humid Period: rapid climate responses to gradual insolation forcing. *Quat. Sci. Rev.* **19**, 347–361 (2000).

Acknowledgements

We thank P. Yodzis for his help in improving the clarity of the manuscript. B. Holling played a key role over the past years in stimulating our discussions around the theme of resilience.

Correspondence and requests for materials should be addressed to M.S. (e-mail: marten.scheffer@aqec.wkao.wau.nl).

Thin-film thermoelectric devices with high room-temperature figures of merit

Rama Venkatasubramanian, Edward Siivola, Thomas Colpitts & Brooks O'Quinn

Research Triangle Institute, Research Triangle Park, North Carolina 27709, USA

Thermoelectric materials are of interest for applications as heat pumps and power generators. The performance of thermoelectric devices is quantified by a figure of merit, ZT , where Z is a measure of a material's thermoelectric properties and T is the absolute temperature. A material with a figure of merit of around unity was first reported over four decades ago, but since then—despite investigation of various approaches—there has been only modest progress in finding materials with enhanced ZT values at room temperature. Here we report thin-film thermoelectric materials that demonstrate a significant enhancement in ZT at 300 K, compared to state-of-the-art bulk Bi_2Te_3 alloys. This amounts to a maximum observed factor of ~ 2.4 for our p-type $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice devices. The enhancement is achieved by controlling the transport of phonons and electrons in the superlattices. Preliminary devices exhibit significant cooling (32 K at around room temperature) and the potential to pump a heat flux of up to 700 W cm^{-2} ; the localized cooling and heating occurs some 23,000 times faster than in bulk devices. We anticipate that the combination of performance, power density and speed achieved in these materials will lead to diverse technological applications: for example, in thermochemistry-on-a-chip, DNA microarrays, fibre-optic switches and microelectrothermal systems.

The performance of thermoelectric devices depends on the figure of merit (ZT) of the material, given by

$$ZT = (\alpha^2 T / \rho K_T) \quad (1)$$

where α , T , ρ and K_T are the Seebeck coefficient, absolute temperature, electrical resistivity and total thermal conductivity, respectively. Z , the material coefficient, can be expressed in terms of lattice thermal conductivity (K_L), electronic thermal conductivity (K_e) and carrier mobility (μ), for a given carrier density (p) and the corresponding α , yielding equation (2), below. Here L_0 is the Lorenz number, approximately $1.5 \times 10^{-8} \text{ V}^2 \text{ K}^{-2}$ in non-degenerate semiconductors, q is the electronic charge, and σ is electrical conductivity.

$$Z = \frac{\alpha^2 \sigma}{K_L + K_e} \approx \frac{\alpha^2}{\left(\frac{K_L}{\mu p q}\right) + L_0 T} \quad (2)$$

State-of-the-art devices utilize alloys, typically p- $\text{Bi}_{1-x}\text{Sb}_{2-x}\text{Te}_{3-y}\text{Se}_y$ ($x \approx 0.5$, $y \approx 0.12$) and n- $\text{Bi}_2(\text{Se}_y\text{Te}_{1-y})_3$ ($y \approx 0.05$) for the 200–400 K temperature range. For certain alloys, K_L can be reduced more strongly than μ , leading¹ to enhanced ZT . A ZT of 0.75 at 300 K in p-type $\text{Bi}_x\text{Sb}_{2-x}\text{Te}_3$ ($x \approx 1$) was reported² 40 years ago. Since then, there has been modest progress in the ZT of thermoelectric materials near 300 K. The highest ZT in any bulk thermoelectric material at 300 K appears to be ~ 1.14 for p-type $(\text{Bi}_2\text{Te}_3)_{0.25}(\text{Sb}_2\text{Te}_3)_{0.72}(\text{Sb}_2\text{Se}_3)_{0.03}$ alloy³, although there has been a report⁴ of an estimated $ZT > 2$ at 300 K in $\text{Bi}_{0.5}\text{Sb}_{1.5}\text{Te}_3$ alloy under a hydrostatic pressure of 2 GPa. A ZT value of greater than one in crystals of Bi–Sb alloy under a magnetic field⁵ at ~ 100 K has been reported. There is no apparent thermodynamic upper limit on ZT ; in addition to its engineering importance, values of ZT significantly larger than unity are of interest to transport theory⁶.

Several possible approaches to enhancing ZT have been investigated. In bulk materials, cage-like structures simulating a phonon glass/electron crystal⁷ have been examined, with a view to reducing K_L without a deterioration of the electronic mobilities. A value of $ZT > 1$ in $\text{LaFe}_3\text{CoSb}_{12}$ was reported⁸ at $T > 700$ K, and attributed primarily to reduction of K_L from La-filling⁹. A ZT of ~ 1.35 was reported for the skutterudite $\text{CeFe}_{3.5}\text{Co}_{0.5}\text{Sb}_{12}$ at ~ 900 K (ref. 10) where, although a dramatic reduction in K_L of the filled skutterudite was observed near 300 K, there was apparently less role of Ce-filling at higher temperatures where the enhanced ZT was observed. A ZT

of ~ 0.8 at 225 K was observed¹¹ in CsBi_4Te_6 , significantly higher than in $\text{Bi}_{0.5}\text{Sb}_{1.5}\text{Te}_3$, apparently due to the smaller bandgap of the former compound (M.G. Kanatzidis, personal communication). From these results, we observe the following: (1) A ZT significantly greater than 1 has not been demonstrated at ordinary temperatures (300 K). (2) A one-to-one correlation between lower K_L and enhanced ZT has not been established. (3) More importantly, the concept of individually tailoring the phonon properties without producing a deterioration of electronic transport, thereby enhancing ZT , has not been established.

Thin-film thermoelectric materials¹² offer tremendous scope for ZT enhancement. Three generic approaches have been proposed to date. One involves the use of quantum-confinement effects¹³ to obtain an enhanced density of states near the Fermi energy. Using such effects, a ZT of 0.9 at 300 K and 2.0 at 550 K, using estimated thermal-conductivity values, has been reported¹⁴ in $\text{PbSe}_{0.98}\text{Te}_{0.02}/\text{PbTe}$ quantum-dot structures. The second approach¹⁵ involves phonon-blocking/electron-transmitting superlattices. These structures utilize the acoustic mismatch between the superlattice components to reduce K_L (refs 16, 17), rather than using the conventional alloying approach, thereby potentially eliminating alloy scattering of carriers. The third thin-film approach is based on thermionic effects in heterostructures^{18,19}. Here we report a ZT at 300 K of ~ 2.4 in p-type $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices and a similar, although less dramatic, $ZT \approx 1.4$ in n-type $\text{Bi}_2\text{Te}_3/\text{Bi}_{2.83}\text{Se}_{0.17}$ superlattices. These ZT values were measured in devices using the Harman technique²⁰, in which parameters related to ρ , K_T and α (that make up ZT) are measured at the same place, at the same time, with current flowing. This method has been extended with a variable-thickness approach (C.B. Vining, personal communication) to obtain the intrinsic ZT and other parameters.

Potential ideality of superlattices in the Bi_2Te_3 system

High-quality superlattices have been produced^{21,22} in the Bi_2Te_3 system, with one of the individual layers as small as $10 \text{ \AA}$, using a low-temperature growth process. Such ultra-short-period superlattices offer significantly higher in-plane carrier mobilities (parallel to the superlattice interfaces) than alloys, owing to the near absence of alloy scattering and random interface carrier scattering²³. We show that the enhanced carrier mobilities in monolayer-range superlattices are effective in the cross-plane direction for certain

superlattices, where we can also obtain reduced K_L and so an enhanced ZT as per equation (2).

We adapted the transmission line model (TLM) technique²³ used for the measurement of specific electrical contact resistivities (ρ_c)²⁴ to determine the cross-plane electrical transport. This adaptation was feasible with the low ρ_c ($< 10^{-7} \Omega \text{ cm}^2$) achievable in these materials. With such ρ_c , the transfer lengths (L_t) are small, and so an effective ‘squeezing’ of the current occurs. A typical TLM device consists of adjacent metallized areas, spaced a few micrometres apart, serving as top current-injection contacts. The current is confined to the thin-film on the substrate. The film between the contacts is etched anisotropically to create various device thickness underneath the contacts²³. Measuring the intercept resistances using successive etch-steps gives the electrical resistivity perpendicular to the superlattice interfaces. We observe that the average length of the contact is $\sim 250 \mu\text{m}$, much greater than L_t ($\sim 1 \mu\text{m}$). Thus the top metal contact, the thermally conducting substrate, and the inactive (where current is not flowing) thermoelectric material act like thermal shunts, in addition to exacerbated radiative/convective losses from the close-to-surface L_t region, for any cooling developed across this region. Thus we have performed an isothermal electrical resistivity measurement, confirmed by negligible Peltier voltage (in a Harman method²⁰ set-up) in such TLM structures. Further, the cross-plane electrical resistivities determined by the TLM method were confirmed by the variable-thickness ZT method in some of the high- ZT samples, as discussed subsequently. This method has been tested on n- Bi_2Te_3 thin films with known theoretical and experimental electrical anisotropy²⁵; we measured an anisotropy of 3.8 ± 0.32 and the theoretical anisotropy is ~ 4.1 (ref. 25). The validity was also confirmed in p-type $\text{Bi}_x\text{Sb}_{2-x}\text{Te}_3$ ($x \approx 0.63 \pm 0.12$) alloy²⁶ films.

The anticipated heterostructure band diagram in these short-period/shallow potential superlattices is shown in Fig. 1a, where the valence-band offset is expected to be less than the average thermal energy of carriers. The variation of the mobility anisotropy as a function of superlattice period, for $d_{\text{Bi}_2\text{Te}_3} = d_{\text{Sb}_2\text{Te}_3}$, is shown in Fig. 1b. $d_{\text{Bi}_2\text{Te}_3}$ and $d_{\text{Sb}_2\text{Te}_3}$ are the thickness of the Bi_2Te_3 and Sb_2Te_3 layers, respectively, that make up a period. The carrier concentration is isotropic, so the ratio of electrical resistivity anisotropy is inversely related to the mobility anisotropy (or $\mu_{\perp}/\mu_{\text{in-plane}}$, where $\mu_{\perp}$ is the cross-plane mobility). The mobility anisotropies of the $10\text{\AA}/10\text{\AA}$, $10\text{\AA}/20\text{\AA}$, $10\text{\AA}/50\text{\AA}$, $20\text{\AA}/20\text{\AA}$, $20\text{\AA}/30\text{\AA}$ and $20\text{\AA}/40\text{\AA}$ structures are shown in Fig. 1c. (Here $10\text{\AA}/50\text{\AA}$, for example, indicates $d_{\text{Bi}_2\text{Te}_3} = 10\text{\AA}$ and $d_{\text{Sb}_2\text{Te}_3} = 50\text{\AA}$ in the superlattice.) The measured cross-plane mobilities in the $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices appear to be consistent with a miniband conduction across the superlattice interfaces, with fully developed or larger valence-band offsets in structures for Bi_2Te_3 thickness $\geq 30\text{\AA}$. The data are qualitatively similar to the transport models developed for $\text{AlGaAs}/\text{GaAs}$ superlattices^{27,28}. Note from Fig. 1b, c that the $10\text{\AA}/50\text{\AA}$, $20\text{\AA}/40\text{\AA}$ and $10\text{\AA}/20\text{\AA}$ $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices offer cross-plane electrical conductivities comparable to in-plane values. It has been shown²⁹ that periods of $50\text{--}60\text{\AA}$ are desirable for minimizing K_L ; thus the $10\text{\AA}/50\text{\AA}$ and $20\text{\AA}/40\text{\AA}$ superlattices are useful for enhanced ZT .

$\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices show significantly reduced K_L . A full description of the measurement procedure, the data, and the mechanisms are presented elsewhere²⁹. The behaviour of these superlattices are consistent with phonon-transmission experiments in $\text{GaAs}/\text{AlGaAs}$ superlattices³⁰, although we have invoked coherent backscattering of phonons at superlattice interfaces²⁹, similar to photon-localization in highly scattering media^{31,32}. Values of K_L of the superlattice structures show a minimum of $\sim 0.22 \text{ W m}^{-1} \text{ K}^{-1}$ at $\sim 50\text{\AA}$; this is a factor of 2.2 smaller than the K_L of $\text{Bi}_{0.5}\text{Sb}_{1.5}\text{Te}_3$ alloy ($\sim 0.49 \text{ W m}^{-1} \text{ K}^{-1}$) along the same c axis. We note that a theory for the thermal conductivity minimum has been proposed³³. K_L for all the $60\text{-\AA}$ -period structures, whether it be $30\text{\AA}/30\text{\AA}$ or $10\text{\AA}/50\text{\AA}$ or

$20\text{\AA}/40\text{\AA}$, is about $0.25 \text{ W m}^{-1} \text{ K}^{-1}$. K_L appears to be more dependent on the superlattice period²⁹, relatively independent of the thickness of the constituents, whereas the electronic transport depends on the period and the relative thickness of the constituents (Fig. 1b, c).

A K_L value of $0.22 \text{ W m}^{-1} \text{ K}^{-1}$ can be compared with the minimum thermal conductivities (K_{min}) predicted for Bi_2Te_3 , using the full-wavelength or half-wavelength models of Slack³⁴ and Cahill³⁵ for the phonon mean free path. Table 1 shows these predicted values, along with experimental K_L values of alloys along the trigonal axis (c axis) or perpendicular to it (a - b axis). The K_L values in the superlattices approach the values of K_{min} estimated from the models of Slack and Cahill. When all the phonons have a mean free path equal to the lattice spacing, K_{min} is expected³⁶ to be between the predictions of the models of Slack and Cahill. The phonon mean free path from kinetic theory is approximately $2.2\text{\AA}$, using a K_L of $0.22 \text{ W m}^{-1} \text{ K}^{-1}$ (ref. 29), assuming that all phonons have similar mean free paths—this is justifiable, given that the Debye temperature of Bi_2Te_3 is $\sim 160\text{ K}$, far lower than the temperature of measurement (300 K).

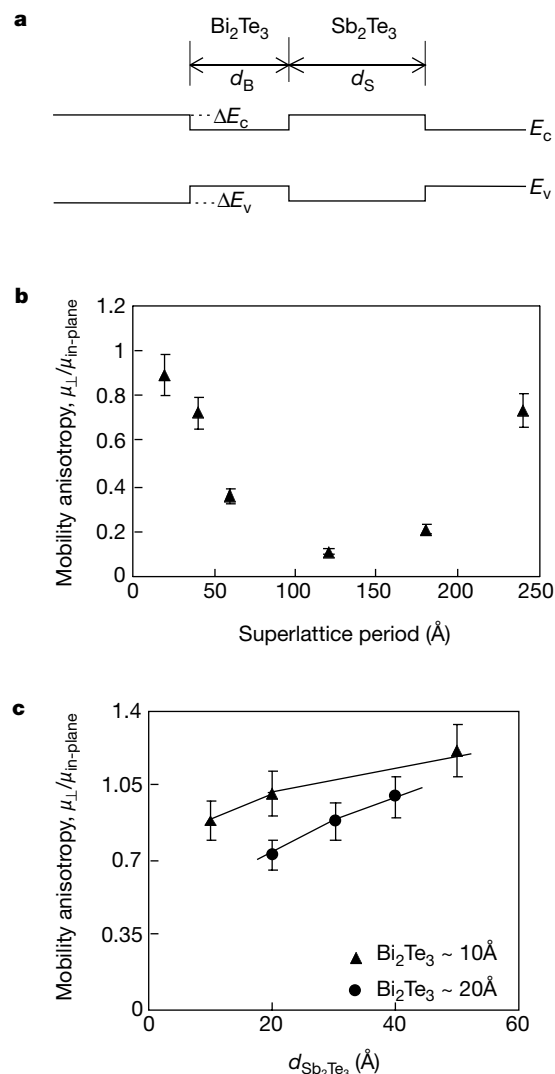


Figure 1 Hole transport across the $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice interface. **a**, The anticipated heterojunction band diagram. $\Delta E_v + \Delta E_c < 2kT$, $\Delta E_v \approx kT$, $d_S + d_B =$ superlattice period (here ΔE_v and ΔE_c represent the valence and conduction band offsets, respectively, in the heterostructure, and $kT \approx 0.0285 \text{ eV}$ at 300 K). **b**, The observed hole mobility anisotropy versus superlattice period for the case of $d_B = d_S = (1/2)$ period. **c**, The hole mobility anisotropy in other $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices.

Table 1 Theoretical and experimental lattice thermal conductivities

Material	Thermal conductivity (W m ⁻¹ K ⁻¹)
$K_{\min}$ of Bi ₂ Te ₃ (a-b axis), Slack model ⁸⁴	0.55
$K_{\min}$ of Bi ₂ Te ₃ (c axis), Slack model ⁸⁴	0.28
$K_{\min}$ of Bi ₂ Te ₃ (a-b axis), Cahill model ⁸⁵	0.28
$K_{\min}$ of Bi ₂ Te ₃ (c axis), Cahill model ⁸⁵	0.14
K_L of Bi _{2-x} Sb _x Te ₃ alloy (a-b axis)	0.97
K_L of Bi _{2-x} Sb _x Te ₃ alloy (c axis)	0.49
K_L of Bi ₂ Te ₃ /Sb ₂ Te ₃ superlattice (c axis)	0.22

Lattice thermal conductivity (K_L) of the Bi₂Te₃/Sb₂Te₃ superlattice (period ~50 Å) compared with K_L observed in the respective alloys and the theoretical minimum lattice thermal conductivity ($K_{\min}$) from various models.

Variable-thickness ZT measurements

ZT of a thermoelectric material can be obtained from a unipolar thermoelement across which a temperature difference is developed by the Peltier effect using a quasi-steady-state current²⁰. The unipolar thermoelement consists of a mesa device with two ohmic contacts for current injection. One end of the device is effectively attached to a heat sink at a fixed temperature by being on a thermally conducting substrate. A schematic of the four-probe measurement is shown in Fig. 2a. A current (I) leads to ohmic (V_R) and Peltier (V_0) voltages across the thermoelement, with the total voltage V_T being $V_R + V_0$ at steady state. With current off, V_R decays within the dielectric relaxation time but V_0 decays according to the thermal time constant. V_R includes resistance from the bulk resistance of the mesa and from that of the two contacts (equation (3)).

$$V_R = \rho_{\perp}(l/a)I + 2(\rho_c/a)I \quad (3)$$

Here, $\rho_{\perp}$, ρ_c (in $\Omega \text{ cm}^2$), l and a are respectively the cross-plane electrical resistivity of the superlattice, the average specific resistivity

of the two contacts, and the height and the cross-sectional area of the device. V_0 , given by equation (4) below, is affected by any difference in heating at the two contacts. As shown below, the ρ_c in the devices is rather small; hence, differences in the two contacts, causing any relative voltages, is even smaller. Even so, currents were reversed, and an average V_0 was used to account for minor differences.

$$V_0 = (\alpha_{\perp}^2/K_T)T(l/a)I \quad (4)$$

From equations (3) and (4), we get extrinsic or device figure of merit (ZT_e) as (V_0/V_R) shown in equation (5)²⁰ below. In thick (large- l) bulk elements, where $2(\rho_c/\rho_{\perp}l)$ is rather small, ZT_e approaches the intrinsic figure of merit (ZT_i) given by equation (1).

$$ZT_e = V_0/V_R = \{\alpha_{\perp}^2 T/K_T \rho_{\perp}\} [1 + 2(\rho_c/\rho_{\perp}l)]^{-1} \quad (5)$$

We can get $\rho_{\perp}$ and $(\alpha_{\perp}^2/K_T)T$, respectively, from the variation of V_R and V_0 with l (or mathematically equivalent, l/a). Thus ZT_i is:

$$ZT_i = [\partial V_0/\partial(l/a)]/[\partial V_R/\partial(l/a)] \quad (6)$$

The data from variable-device-thickness experiments are shown in Fig. 2b and c. We point out that each data point in Fig. 2b and c is an average of 125 transient measurements, stored digitally and averaged out. The variable-thickness method for extracting ZT_i works for medium-to-good ρ_c , that is, medium-to-good device ZT_e shown in Fig. 2b and c, respectively. In Fig. 2b, we obtain ZT_e (given by V_0/V_R) for the largest mesa thickness of ~1.35 μm of ~0.66; even so, the ratio of the two slopes, from equation (6) above, gives a ZT_i of 2.07. The y -intercept of V_R versus l/a is used to obtain ρ_c , ~1.45 $\times 10^{-7} \Omega \text{ cm}^2$ for the sample in Fig. 2b. This is more than a factor of ten smaller than in bulk technology, but still high for thin-film devices, limiting the extrinsic ZT to ~0.66. In Fig. 2c, however, a combination of thicker 10 Å/50 Å Bi₂Te₃/Sb₂Te₃ superlattice (thickest mesa ~2.67 μm) and an improved ρ_c of

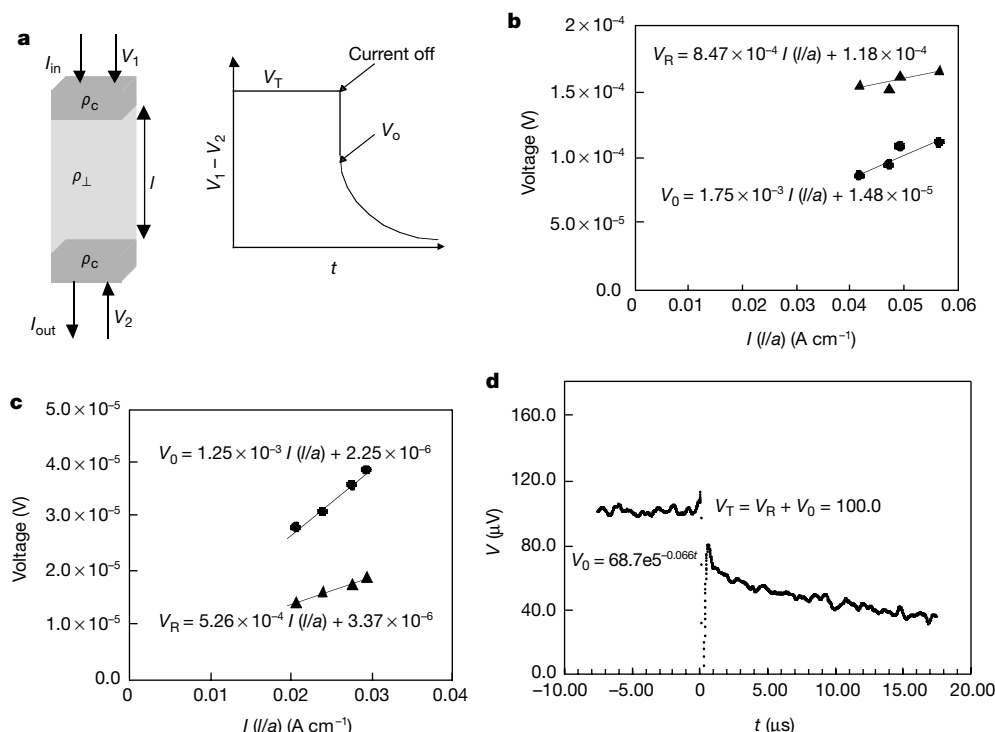


Figure 2 The ZT of a thermoelectric device. **a**, Direct measurement by a four-probe Harman method. (See text for definitions of symbols.) **b**, **c**, Variable-thickness ZT measurements. V_R and V_0 obtained from Harman method for each thickness indicate $ZT_i \approx 2.07$ and 2.34 in **b** and **c**, respectively. Working conditions: **b**, $I = 20.5$ mA and area, $4.9 \times 10^{-5} \text{ cm}^2$; **c**, $I = 20.3$ mA and area, $1.8 \times 10^{-4} \text{ cm}^2$. Triangles, V_R ; circles, V_0 . **d**, Harman-method transient on a 5.4- μm -thick 10 Å/50 Å Bi₂Te₃/Sb₂Te₃ superlattice

device, shown for one current direction. The average extrinsic ZT for the two current directions is 2.38 ± 0.19 . The voltage spike at $t \approx 0$, from momentary inductive coupling between the current and voltage circuits, prevents reliable determination of residual V_0 for $t < 1 \mu\text{s}$; thus extrapolation of data is used from $t > 1 \mu\text{s}$ to get V_0 . The transient is an average of 125 consecutive measurements, stored digitally and averaged out, which should lead to minimal error in extrapolation in spite of the spike.

$1.5 \times 10^{-8} \Omega \text{ cm}^2$ leads to a ZT_c value of ~ 2.04 for the thickest device. Based on the ratio of two slopes, a ZT_i of ~ 2.34 is deduced.

For the $10\text{\AA}/50\text{\AA}$ $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice in Fig. 2b, the measured $\rho_{\perp}$ is $8.47 \times 10^{-4} \Omega \text{ cm}$. With $\rho_{\text{in-plane}}$ of $9.48 \times 10^{-4} \Omega \text{ cm}$, $\rho_{\perp}/\rho_{\text{in-plane}}$ or $\mu_{\perp}/\mu_{\text{in-plane}}$ is ~ 1.12 . For the sample in Fig. 2c, $\rho_{\perp}$ is $5.26 \times 10^{-4} \Omega \text{ cm}$ and $\rho_{\text{in-plane}}$ is $5.5 \times 10^{-4} \Omega \text{ cm}$, thus giving an anisotropy of 1.05. These values are comparable to that indicated in Fig. 1c for the $10\text{\AA}/50\text{\AA}$ $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice, confirming the ideal electron (hole) transmission characteristics of the structure.

Figure 2d shows the Harman-method transient²⁰ for a $\sim 5.4\text{-}\mu\text{m}$ -thick $10\text{\AA}/50\text{\AA}$ $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice thermoelement. The average ZT_c for two current directions was 2.38 ± 0.19 at 300 K. The material parameters for this superlattice film indicated a ZT_i of 2.59, and the measured ZT_c of 2.38 translates to a ρ_c of $\sim 1.3 \times 10^{-8} \Omega \text{ cm}^2$. Thus, going from a superlattice device of $\sim 1.35 \mu\text{m}$ to $\sim 2.67 \mu\text{m}$ to $\sim 5.4 \mu\text{m}$, we see no significant change in ZT_i and so the ZT_c begins to approach ZT_i because of a decreasing role of ρ_c .

As a check, we have measured ZT on 1-mm-thick and thinned-down ($5\text{--}20 \mu\text{m}$) thermoelements made from bulk p-type $\text{Bi}_{1-x}\text{Sb}_{2-x}\text{Te}_{3-y}\text{Se}_y$ ($x \approx 0.63 \pm 0.12$; $y \approx 0.12$) alloys, comparable to those studied in ref. 3. The variable-thickness method indicates a ZT_i of ~ 1.09 at 300 K along the a - b axis in agreement with the results of ref. 3, indicating no unexpected benefits resulting from utilizing a thin-film version of the commercial material. Also, ZT in alloy films (non-superlattice structures) along the c axis were in the range of 0.4 ± 0.13 for carrier levels of $\sim 3 \times 10^{19} \text{ cm}^{-3}$, consistent with reported values for bulk materials²⁶.

Using temperature-dependent measurements of ρ , α and K_T of the p-type superlattice—and noting that there is little electrical anisotropy—the estimated ZT_i versus T of $10\text{\AA}/50\text{\AA}$ p-type $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice is shown in Fig. 3, along with those of other materials. It appears that the superlattices would offer enhanced performance compared to the bulk p-type $\text{Bi}_{2-x}\text{Sb}_x\text{Te}_3$ alloys at lower temperatures. This is not surprising, given that the efficacy of superlattices in reducing K_L improves at lower temperatures as in Si/Ge superlattices¹⁷. As ‘proof of advantage’ of enhanced ZT at lower temperatures, we have obtained four times the cooling with $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice devices compared to bulk p-type $\text{Bi}_{2-x}\text{Sb}_x\text{Te}_3$ devices at 210 K, for similar l/a , thermal load and parasitics. (Parasitics are unintentional heat losses from conductive, convective and any radiative processes.) Figure 3 shows that the p-type $10\text{\AA}/50\text{\AA}$ $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice offers improved ZT compared to CsBi_4Te_6 alloy¹¹ at ~ 210 K.

Phonon-blocking/electron-transmitting structures

The results obtained with the $10\text{\AA}/50\text{\AA}$ $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices indicate that we can fine-tune the phonon and hole (charge carriers)

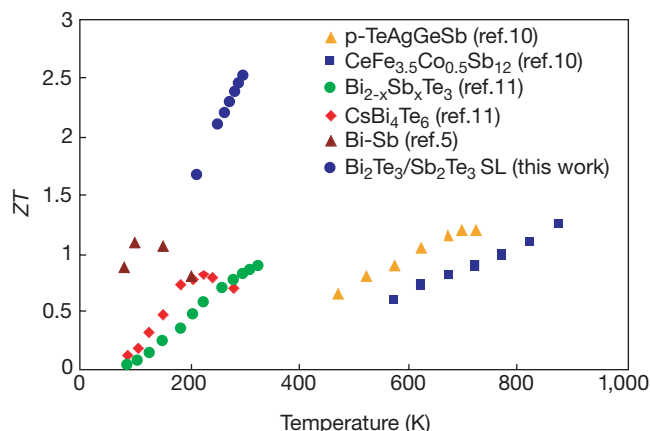


Figure 3 Temperature dependence of ZT of $10\text{\AA}/50\text{\AA}$ p-type $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattice compared to those of several recently reported materials.

transport³⁷ to improve ZT . Other structures that show $ZT > 1$ at 300 K include $10\text{\AA}/40\text{\AA}$ and $20\text{\AA}/40\text{\AA}$ $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices. It is useful to compare the transport characteristics of phonons and holes. The product³⁸ kl_{mfp} , where k and l_{mfp} are respectively the average wavevector and mean free path, is used as a measure of the amount of blocking (or lack of it) for phonons and holes. For holes, we estimate the thermal velocity (v_{th}), de Broglie wavelength, and the wavevector magnitude as $\sim 2.1 \times 10^7 \text{ cm s}^{-1}$, $\sim 114 \text{\AA}$, and $\sim 5.5 \times 10^6 \text{ cm}^{-1}$, respectively. From the relation³⁹ $(l_{\text{mfp}})_{\perp} \approx (v_{\text{th}}/m^* \mu_{\perp}/q)$, we obtain $(l_{\text{mfp}})_{\perp}$ as $136 \text{\AA}$ (for the sample with $ZT_i \approx 2.34$, $\mu_{\perp}$ of $383 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ is deduced from $\rho_{\perp}$, obtained from variable-thickness ZT). (Here m^* is the effective mass of charge carriers; holes in this case.) For an average phonon wavelength of $\sim 30 \text{\AA}$ and hole wavelength of $\sim 114 \text{\AA}$, we obtain $(kl_{\text{mfp}})_{\text{phonons}} \approx 0.5$ and $(kl_{\text{mfp}})_{\text{holes}} \approx 7.6$. We believe that this comparison probably captures the phonon-blocking/electron-transmitting nature of certain superlattices³⁷.

n-type superlattices

We have also obtained encouraging results with n-type $10\text{\AA}/50\text{\AA}$ $\text{Bi}_2\text{Te}_3/\text{Bi}_2\text{Te}_{2.83}\text{Se}_{0.17}$ superlattices, indicating $ZT_i > 1$ at 300 K. The variable-thickness ZT measurements, similar to those shown in Fig. 2b and c, indicate a ZT_i of ~ 1.46 at 300 K in these superlattices. The best extrinsic ZT_c for an n-type device has been ~ 1.2 at 300 K. The closeness of ZT_i and ZT_c is from obtaining a ρ_c value of $\sim 1.2 \times 10^{-8} \Omega \text{ cm}^2$. The cross-plane (along c axis) electrical resistivity in these n-type superlattices is $\sim 1.23 \times 10^{-3} \Omega \text{ cm}$, not significantly higher than the in-plane (along a - b axis) electrical resistivity of $1.04 \times 10^{-3} \Omega \text{ cm}$. Thus in the $10\text{\AA}/50\text{\AA}$ n-type superlattices, apparently due to weak-confinement/near-zero band-offset, there is minimal anisotropy between in-plane and cross-plane electrical resistivities, similar to the $10\text{\AA}/50\text{\AA}$ and other p-type $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices (Fig. 1c). The lack of electrical anisotropy in both the high-performance p-type and n-type superlattices, comparing properties along the a - b and c crystallographic axes, is in marked contrast to the electrical anisotropy observed in both p-type and n-type bulk materials²⁶.

The reason for the less-than-impressive ZT_i of 1.46 at 300 K in $\text{Bi}_2\text{Te}_3/\text{Bi}_2\text{Te}_{2.83}\text{Se}_{0.17}$ superlattices, compared to ~ 2.4 at 300 K in the best p-type $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices, is its higher K_L . From its $\partial V_0/\partial(I/l/a)$ and an α of $\sim 238 \mu\text{V K}^{-1}$, a cross-plane K_T of $\sim 9.45 \text{ mW cm}^{-1} \text{ K}^{-1}$ is obtained. Using the Weidemann–Franz law and a $\rho_{\perp}$ value of $1.23 \times 10^{-3} \Omega \text{ cm}$, we obtain an electronic thermal conductivity of $\sim 3.7 \text{ mW cm}^{-1} \text{ K}^{-1}$. Thus the cross-plane lattice thermal conductivity is $\sim 5.8 \text{ mW cm}^{-1} \text{ K}^{-1}$. This is more like the c -axis K_L of bulk alloys²⁶—and much higher than the value of $\sim 2.5 \text{ mW cm}^{-1} \text{ K}^{-1}$ observed in $60\text{-}\text{\AA}$ -period p-type $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices in ref. 29, and higher than that derived from the analysis of data in Figs. 2b, c. We believe that these results indicate near-ideal superlattice interfaces in the $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ system²¹, where the compositional modifications of either Bi or Sb are accomplished within regions enclosed by the Te–Te van der Waals bond. Thus, it should be possible to obtain mirror-like superlattice interfaces, leading to potential reflection effects for reducing K_L (ref. 29). In contrast, in the $\text{Bi}_2\text{Te}_3/\text{Bi}_2\text{Te}_{2.83}\text{Se}_{0.17}$ system, where both Se and Te are expected to be present at the van der Waals interface, we anticipate much unintended compositional mixing. Thus, the lattice thermal conductivity is likely to be more typical of an alloy. This was further evident in our observations of a lack of any significant improvement in electronic mobilities in n-type $\text{Bi}_2\text{Te}_3/\text{Bi}_2\text{Te}_{2.83}\text{Se}_{0.17}$ superlattices (compared to n-type $\text{Bi}_2\text{Te}_{3-x}\text{Se}_x$ alloys), in contrast to a marked enhancement of carrier mobilities observed in p-type $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ superlattices relative to p-type $\text{Bi}_{2-x}\text{Sb}_x\text{Te}_3$ alloys²¹. Therefore, enhancement of ZT_i to 1.46 in n-type superlattices (from ~ 1 for bulk n-type alloy) appears to be attributable to the lack of electrical anisotropy and the typical lower lattice thermal conductivity associated with the c axis ($\sim 5.8 \text{ mW cm}^{-1} \text{ K}^{-1}$

compared to $\sim 10 \text{ mW cm}^{-1} \text{ K}^{-1}$ along the a - b axis). Better n-type superlattices with low K_L could be possible in those superlattices that have negligible compositional mixing at the van der Waals interface.

Thermoelectric devices for localized, rapid cooling/heating

Figure 4a shows the sub-ambient cooling obtained in a p-type superlattice device. With an ambient temperature T_{ambient} of about 298 K, we have 32.2 K of cooling (down to -7°C)—without any forced heat removal by blowing air or running water at the heat sink. Cooling was measured with $12.5\text{-}\mu\text{m}$ thermocouple wires (bead size $\sim 50\text{-}\mu\text{m}$). Thus the measured cooling is almost always with a load comparable to the thin-film device and in spite of the significant probe-heating on top of the mesa. Under similar testing conditions (same l/a , lack of heat removal at heat sink, residual

thermal load/heat loss), we measured $\sim 18.4 \text{ K}$ of cooling in p-type bulk devices (Fig. 4a). In the linear regime (current $< 1 \text{ A}$) where heat-sinking considerations are less, the superlattice device shows a factor of 2.2 cooling compared to bulk. We observed 40 K of cooling for the superlattice device when the heat sink is maintained at 353 K using a large-wattage heater, equivalent to a semi-infinite heat-source. This is analogous to the cooling of heat-sensitive devices in regions adjacent to hotspots in a chip. These results can be compared to recent measurements of cooling of 2.8 K and 6.9 K, measured using micro-thermocouples at 298 K and 373 K, respectively, in SiGeC/Si devices⁴⁰ of larger l/a ratio (~ 12.5).

Thin-film thermoelements lead to large cooling power densities (P_D) for similar maximum heat pumping (Q_m)⁴¹ as per equation (7) below, following equation (1). As the area of the device, a , is reduced proportional to its thickness, l , Q_m is unaffected. However, cooling power density ($P_D \approx Q/a$) is increased. Here $T_{\text{hot-side}}$ and $T_{\text{cold-side}}$ refer to the absolute temperature of the hot and cold junctions, respectively, at either end of the active thermoelement.

$$Q_m = (al)\{[0.5\alpha^2 T_{\text{cold-side}}^2 / \rho] - [K_T(T_{\text{hot-side}} - T_{\text{cold-side}})]\} \quad (7)$$

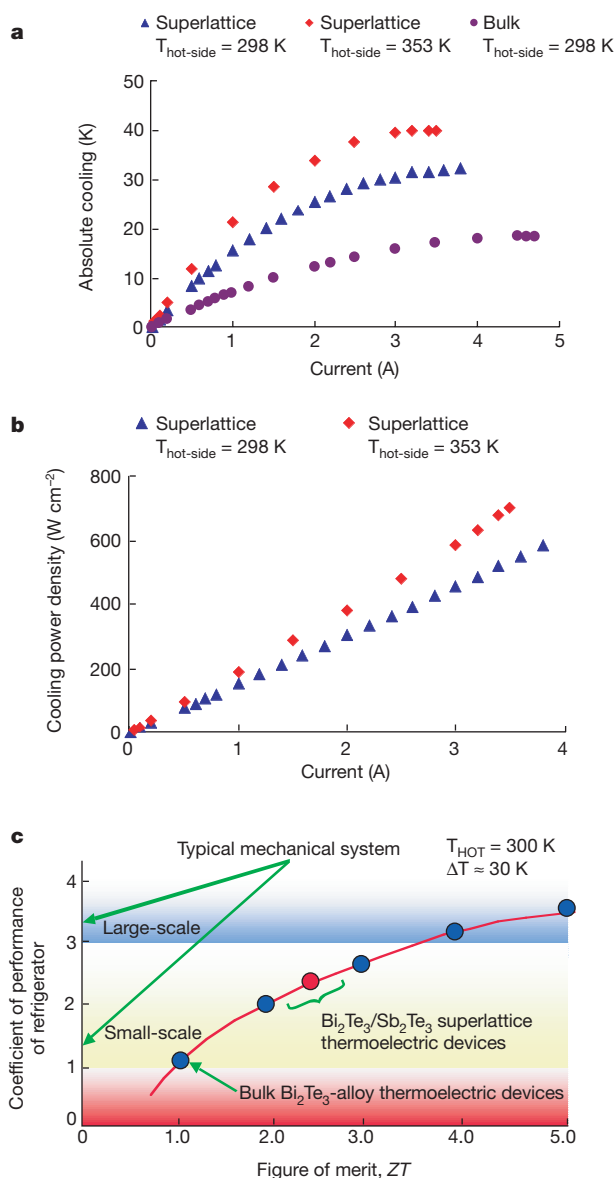


Figure 4 Cooling properties. **a**, Observed cooling as a function of current in a $150\text{-}\mu\text{m}$ circular device made from a p-type superlattice (data shown for hot side at 298 K (blue triangles) and 353 K (red diamonds)) is compared with that of a bulk device (hot side at 298 K, purple circles). **b**, Estimated cooling power density for superlattice devices as a function of current; blue triangles, hot side at 298 K, red diamonds, hot side at 353 K. **c**, Potential COP as a function of ZT with various technologies. T_{HOT} refers to the heat-sink temperature.

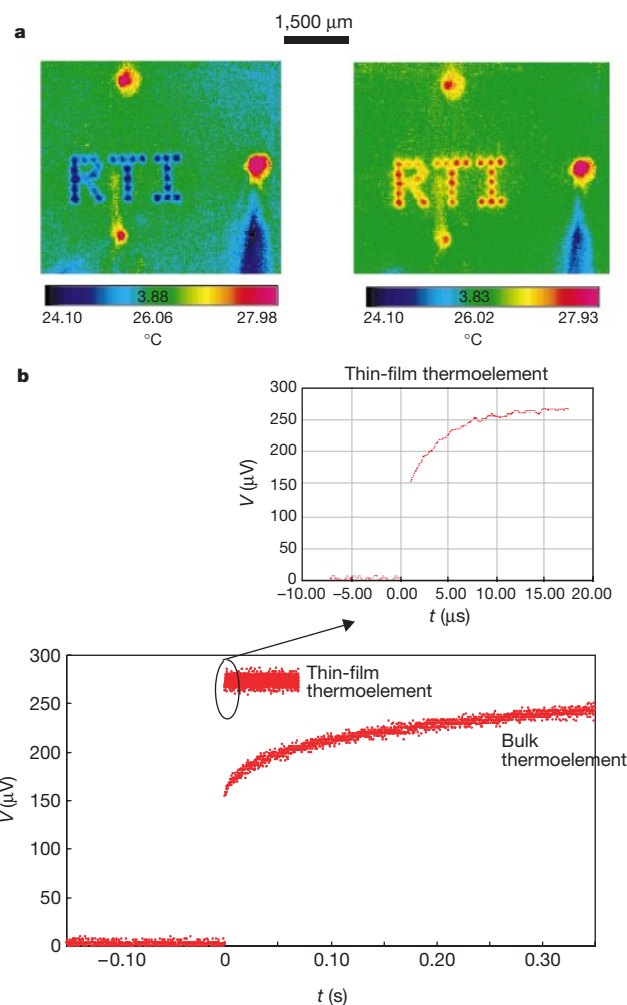


Figure 5 Demonstration of localized and high-speed cooling/heating obtainable with thin-film devices. **a**, Infrared images of spot cooling/heating with thin-film (p-type only) superlattice devices. Note current-injection areas (probes) do not change with current reversal; current through each element, $\sim 65 \text{ mA}$. The infrared images of spots, corresponding to localized cooling or heating, spelling out RTI (Research Triangle Institute) were made using photolithographically processed micro-thermoelements. **b**, Comparison of thermal/cooling time response of thin-film ($\sim 5\text{-}\mu\text{m}$) superlattice device and a bulk device.

In addition to offering low ZT , which translates to low coefficient of performance (COP) in refrigeration, the bulk thermoelectric devices have low P_D ($\sim 1 \text{ W cm}^{-2}$), another disincentive for high-power electronic/microprocessor applications. Using the values of K_T and α of the superlattices, available P_D as a function of current is estimated in Fig. 4b for the superlattice device. We estimate a value of P_D of $\sim 700 \text{ W cm}^{-2}$ at 353 K and $\sim 585 \text{ W cm}^{-2}$ at 298 K at the measured maximum cooling in superlattice devices compared to a value of $\sim 1.9 \text{ W cm}^{-2}$ in the bulk device of Fig. 4a. The effect of higher ZT on COP is shown in Fig. 4c. The translation of the device ZT to COP of refrigeration would involve efficient heat removal at the heat sink, reduction of thermal resistances at the interface between the active device and the two heat-spreaders, and the fabrication of p–n couples with minimal interconnect resistances. A sub-ambient cooling of 22.5 K at 300 K has been obtained in early p–n couples, currently limited by lower ZT of the n-thermoelement, large interconnect-resistance between the p- and n-thermoelements and absence of forced heat removal.

The thin-film devices also allow the concept of localized cooling by matching the footprint of the refrigeration devices to that of thermal load. This is made feasible by the combination of micro-electronic processing⁴² and the ability to place the thermoelectric devices at points of interest; the reversal of current allows spot heating. This flexibility is demonstrated by the infrared images in Fig. 5a. In addition to placing cooling power where required, these thin-film superlattice thermoelements are fast-acting—about 23,000 times faster than bulk devices. This faster response is demonstrated in Fig. 5b. The thin-film device achieves steady-state cooling (indicated by the development of cooling-induced Seebeck voltage) in 15 μs , while the bulk thermoelement takes about 0.35 s. This is a result of the response time associated with the transport of heat through the thin film (micrometres) rather than through the millimetres associated with bulk devices. The thermal response time is about $4l^2/\pi^2 D$, where l is the thickness of the thermoelement and D is the thermal diffusivity. This rapid, high-performance cooling and heating, capable of high power densities and with the ability to be locally applied, could have applications in technologies ranging from genomic/proteomic chips to fibre-optic switching³⁷.

Power conversion with thin-film superlattice devices

We also tested the thin-film microdevices in a power-conversion mode. These superlattice thin-film devices, 5.2 μm thick, can develop a ΔT of 70 K across them, with a corresponding open-circuit voltage indicating an average α of $\sim 243 \mu\text{V K}^{-1}$. The ΔT translates into a temperature gradient of $\sim 134,000 \text{ K cm}^{-1}$. The typical Seebeck coefficient suggests no unusual departure from classical behaviour at these gradients. Due to the higher ZT , modest power-conversion efficiencies can be achieved even for low ΔT . Lightweight, high-power-density devices could be useful for portable power and thermal scavenging.

Thin-film thermoelectric technology could be implemented in a modular fashion: that is, units of thin-film thermoelectric heat-pump or power-conversion devices could be scaled for various capacities. We have reported here several techniques that could be useful in creating such modules: the bonding of thin films to a heat sink, the removal of substrate, photo-lithography to achieve appropriate l/a ratios, the preparation of low-resistivity contacts, and the ability to preserve the properties of the nanometre-scale superlattices during detailed processing. □

Received 6 June; accepted 24 August 2001.

1. Ioffe, A. F. *Semiconductor Thermoelements and Thermoelectric Cooling* (Infosearch, London, 1957).
2. Wright, D. A. Thermoelectric properties of bismuth telluride and its alloys. *Nature* **181**, 834 (1958).
3. Ettenberg, M. H., Jessor, W. A. & Rosi, F. D. in *Proc. 15th Int. Conf. on Thermoelectrics* (ed. Caillat, T.) 52–56 (IEEE, Piscataway, NJ, 1996).
4. Polvani, D. A., Meng, J. F., Chandrasekar, N. V., Sharp, J. & Badding, J. V. Large improvement in

- thermoelectric properties in pressure-tuned p-type $\text{Sb}_{1.5}\text{Bi}_{0.5}\text{Te}_3$. *Chem. Mater.* **13**, 2068–2071 (2001).
5. Yim, W. M. & Amith, A. Bi-Sb alloys for magneto-thermoelectric and thermomagnetic cooling. *Solid State Electron.* **15**, 1141–1165 (1972).
6. Vining, C. B. in *Proc. 11th Int. Conf. on Thermoelectrics* (ed. Rao, K. R.) 276–284 (Univ. Texas at Arlington, 1992).
7. Slack, G. A. & Tsoukala, V. G. Some properties of semiconducting IrSb_3 . *J. Appl. Phys.* **76**, 1665–1671 (1994).
8. Sales, B. C., Mandrus, D. & Williams, R. K. Filled skutterudite antimonides: a new class of thermoelectric materials. *Science* **272**, 1325–1328 (1996).
9. Mandrus, D. *et al.* Filled skutterudite antimonides: Validation of the electron-crystal phonon-glass approach to new thermoelectric materials. *Mater. Res. Soc. Symp. Proc.* **478**, 199–209 (1997).
10. Fleurial, J. P. *et al.* in *Proc. 15th Int. Conf. on Thermoelectrics* (ed. Caillat, T.) 91–95 (IEEE, Piscataway, NJ, 1996).
11. Chung, D. Y. *et al.* CsBi_4Te_6 : A high-performance thermoelectric material for low-temperature application. *Science* **287**, 1024–1027 (2000).
12. Tritt, T. M., Kanatzidis, M. G., Lyon, H. B. Jr & Mahan, G. D. Thermoelectric materials—New directions and approaches. *Mater. Res. Soc. Proc.* **478**, 73–84 (1997).
13. Hicks, L. D. & Dresselhaus, M. D. Effect of quantum-well structures on the thermoelectric figure of merit. *Phys. Rev. B* **47**, 12727–12731 (1993).
14. Harman, T. C., Taylor, P. J., Spears, D. L. & Walsh, M. P. in *Proc. 18th Int. Conf. on Thermoelectrics* (ed. Ehrlich, A.) 280–284 (IEEE, Piscataway, NJ, 1999).
15. Venkatasubramanian, R. *et al.* in *Proc. 1st Natl Thermogenic Cooler Workshop* (ed. Horn, S. B.) 196–231 (Center for Night Vision and Electro-Optics, Fort Belvoir, VA, 1992).
16. Venkatasubramanian, R. Thin-film superlattice and quantum-well structures—a new approach to high-performance thermoelectric materials. *Naval Res. Rev.* **58**, 31–40 (1996).
17. Lee, S. M., Cahill, D. G. & Venkatasubramanian, R. Thermal conductivity of Si-Ge superlattices. *Appl. Phys. Lett.* **70**, 2957–2959 (1997).
18. Mahan, G. D. & Woods, L. M. Multilayer thermionic refrigeration. *Phys. Rev. Lett.* **80**, 4016–4019 (1998).
19. Shakouri, A. & Bowers, J. E. Heterostructure integrated thermionic coolers. *Appl. Phys. Lett.* **71**, 1234–1236 (1997).
20. Harman, T. C., Cahn, J. H. & Logan, M. J. Measurement of thermal conductivity by utilization of the Peltier effect. *J. Appl. Phys.* **30**, 1351–1359 (1959).
21. Venkatasubramanian, R. *et al.* Low-temperature organometallic epitaxy and its application to superlattice structures in thermoelectrics. *Appl. Phys. Lett.* **75**, 1104–1106 (1999).
22. Venkatasubramanian, R. Low temperature chemical vapor deposition and etching apparatus and method. US Patent No. 6071351 (6 June 2000).
23. Venkatasubramanian, R. in *Recent Trends in Thermoelectric Materials Research III* (ed. Tritt, T. M.) Ch. 4 (Academic, San Diego, 2001).
24. Berger, H. H. Models for contacts to planar devices. *Solid State Electron.* **15**, 145–158 (1972).
25. Drabble, J. R., Groves, R. D. & Wolfe, R. Galvanomagnetic effects in n-type Bi_2Te_3 . *Proc. Phys. Soc.* **71**, 430–434 (1957).
26. Scherrer, H. & Scherrer, S. in *CRC Handbook of Thermoelectrics* (ed. Rowe, D. M.) 211–237 (CRC Press, Boca Raton, FL, 1995).
27. Palmier, J. F. in *Heterojunctions and Semiconductor Superlattices* (eds Allan, G., Bastard, G., Boccaro, N., Launo, N. & Voos, M.) 127–145 (Springer, Berlin, 1986).
28. Cappaso, F., Mohammed, K. & Cho, A. Y. Resonant tunneling through double barriers, perpendicular quantum transport phenomena in superlattices, and their device applications. *IEEE J. Quant. Electron.* **QE-22**, 1853–1869 (1986).
29. Venkatasubramanian, R. Lattice thermal conductivity reduction and phonon localizationlike behavior in superlattice structures. *Phys. Rev. B* **61**, 3091–3097 (2000).
30. Narayanaamurti, V., Störmer, H. L., Chin, M. A., Gossard, A. C. & Weigmann, W. Selective transmission of high-frequency phonons by a superlattice: the dielectric phonon filter. *Phys. Rev. Lett.* **43**, 2012–2015 (1979).
31. John, S. Strong localization of photons in certain disordered dielectric superlattices. *Phys. Rev. Lett.* **58**, 2486–2489 (1987).
32. Wiersma, D. S., Bartolini, P., Lagendik, A. & Righini, R. Localization of light in a disordered medium. *Nature* **390**, 671–673 (1997).
33. Simkin, M. V. & Mahan, G. D. Minimum thermal conductivity of superlattices. *Phys. Rev. Lett.* **84**, 927–930 (2000).
34. Slack, G. in *Solid State Physics*, 1–71 (eds Ehrenreich, H., Seitz, F. & Turnbull, D.) Series 34 (Academic, New York, 1979).
35. Cahill, D. G., Watson, S. K. & Pohl, R. O. Lower limit to the thermal conductivity of disordered crystals. *Phys. Rev. B* **46**, 6131–6140 (1992).
36. Goldsmid, H. J. in *Proc. 18th Int. Conf. on Thermoelectrics* (ed. Ehrlich, A.) 531–535 (IEEE Press, Piscataway, NJ, 1999).
37. Venkatasubramanian, R. *Thin-film Thermoelectric Cooling and Heating Devices for DNA Genomics/Proteomics, Thermo-Optical Switching-Circuits, and IR Tags* (US Patent Filing, Ser. No. 60/282,185, 2001).
38. Sheng, P. *Introduction to Wave Scattering, Localization, and Mesoscopic Phenomena* (Academic, London, 1995).
39. Kittel, C. *Introduction to Solid State Physics* (Wiley, New York, 1976).
40. Fan, X. *et al.* SiGeC/Si superlattice microcoolers. *Appl. Phys. Lett.* **78**, 1580–1582 (2001).
41. Cadoff, I. B. & Miller, E. *Thermoelectric Materials and Devices* (Reinhold, New York, 1960).
42. Venkatasubramanian, R. Thin-film thermoelectric device and fabrication method of same. US Patent No. 6300150 (9 October 2001).

Acknowledgements

This work was supported by the US Defense Advanced Research Projects Agency and the US Office of Naval Research.

Correspondence and requests for materials should be addressed to R.V. (e-mail: rama@rti.org).

A sperm ion channel required for sperm motility and male fertility

Dejian Ren*, Betsy Navarro*, Gloria Perez†, Alexander C. Jackson*, Shyuefang Hsu*, Qing Shi*, Jonathan L. Tilly† & David E. Clapham*

* Howard Hughes Medical Institute, Children's Hospital, Harvard Medical School, Enders 1309, 320 Longwood Avenue, Boston, Massachusetts 02115, USA

† Vincent Center for Reproductive Biology, Massachusetts General Hospital and Department of Obstetrics, Gynecology and Reproductive Biology, Harvard Medical School, Boston, Massachusetts 02114, USA

Calcium and cyclic nucleotides have crucial roles in mammalian fertilization, but the molecules comprising the Ca^{2+} -permeation pathway in sperm motility are poorly understood. Here we describe a putative sperm cation channel, CatSper, whose amino-acid sequence most closely resembles a single, six-transmembrane-spanning repeat of the voltage-dependent Ca^{2+} -channel four-repeat structure. CatSper is located specifically in the principal piece of the sperm tail. Targeted disruption of the gene results in male sterility in otherwise normal mice. Sperm motility is decreased markedly in CatSper^{-/-} mice, and CatSper^{-/-} sperm are unable to fertilize intact eggs. In addition, the cyclic-AMP-induced Ca^{2+} influx is abolished in the sperm of mutant mice. CatSper is thus vital to cAMP-mediated Ca^{2+} influx in sperm, sperm motility and fertilization. CatSper represents an excellent target for non-hormonal contraceptives for both men and women.

Male sperm and female eggs interact reciprocally in mammalian fertilization^{1,2}. To reach the site of fertilization, sperm must travel long distances and become primed for fertilization of the eggs through capacitation and other processes. Once they arrive at the surface of the egg, sperm interact with the egg's extracellular matrix glycoproteins including the zona pellucida proteins. Sperm release acidic material during the acrosome reaction, a signalling event that presumably involves the opening of Ca^{2+} channels and the influx of Ca^{2+} into the sperm heads³. Transient receptor potential protein channel 2 (TRPC2), a putative Ca^{2+} -permeant channel, has been implicated in the acrosome reaction⁴. Penetration of sperm through the thick outer layer of the egg is achieved through chemical lysis of the egg coat and/or the mechanical motion of sperm⁵. After infiltrating the egg's zona pellucida coat, the sperm membrane fuses with that of the egg. Fusion is followed by activation of the fertilization process, beginning with Ca^{2+} oscillations in the egg^{1,2}.

Ca^{2+} and cyclic nucleotides control sperm motility^{6–8} and several voltage-dependent Ca^{2+} -channel (Ca_v) messenger RNAs and cyclic-nucleotide-gated (CNG) proteins have been detected in sperm cell precursors^{7,9–11}. Furthermore, low-voltage-activated, dihydropyridine-sensitive "T-type" channels^{12,13} and pharmacologically defined N- and R-type currents have been measured in spermatogenic cells¹⁴. But the role of these channels in spermatogenesis or mature sperm function is not known.

Here we describe the cloning and functional characterization of an unusual sperm cation channel. The CatSper gene is unique in that it encodes a single, six-transmembrane-spanning repeat (like voltage-dependent K^+ (K_v) channels), but its pore region and overall homology is closest to a single domain of the much larger four-repeat Ca_v channels. The gene product is expressed exclusively in the testis and not in other tissues such as the brain, heart, kidney or immune system. In sperm, the channel is localized primarily to the tail's principal piece, not the head or midpiece. As we show, poor sperm motility and male infertility result from gene-targeted elimination of the mouse CatSper protein.

Cloning of the putative one-repeat cation channel CatSper

While searching for Ca^{2+} channels, we discovered a fragment of an expressed sequence tag (EST) complementary DNA (accession number AA416682) that showed similarity to known voltage-gated Ca^{2+} channels. Expression analysis showed that it was present only in testis (see below). Polymerase chain reaction (PCR) and

library screening were used to clone the full-length cDNAs from human and mouse testis. The gene for CatSper (cation channel of sperm) predicted a primary structure of 686 amino acids (Fig. 1a).

In a BLAST search (see <http://www.ncbi.nlm.nih.gov>), the functional proteins with closest similarity to CatSper were voltage-gated Ca^{2+} channels, not K_v , CNG or TRP channels. The pore-forming subunits (α_1) of Ca_v channels have four repeats of a six-transmembrane-spanning domain¹⁵. Located in the putative channel pore region of each of the four repeats are glutamine/aspartate residues

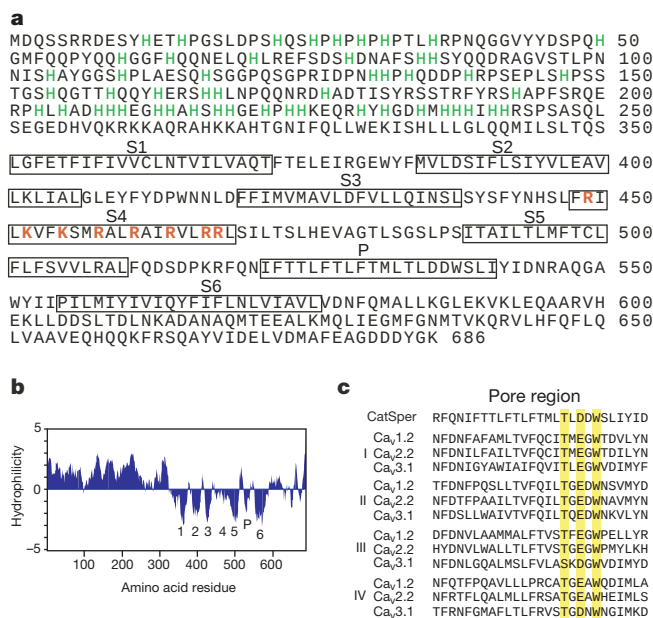


Figure 1 Primary structure of mouse CatSper. **a**, Amino-acid sequence; the six putative transmembrane domains (S1–S6) and the pore (P) region are boxed. The positively charged amino acids (K/R) in the S4 region are shown in red. Histidines in the first 250 amino acids are labelled in green. **b**, Hydropathy plot of CatSper predicts six transmembrane domains (1–6) and a P loop (P). Window size is 11. **c**, Alignment of the putative pore region of CatSper with that of the four domains (I, II, III, IV) from Ca_v 1–3. The conserved residues in the T/S-x-E/D-x-W motif are shaded. GenBank accession numbers: X15539, Ca_v 1.2; M94172, Ca_v 2.2; 054898, Ca_v 3.1; mouse CatSper, AF407332; human CatSper, AF407333.

that impart Ca^{2+} selectivity on the channel by coordinating Ca^{2+} ions¹⁶. These residues are conserved among all the established Ca_v channels and CatSper (Fig. 1c).

As is characteristic for voltage-gated channels, positively charged amino acids (lysine/arginine) are interspersed every three amino acids in the transmembrane region of the putative CatSper S4 domain. Like K_v , CNG and TRP channels, a hydrophobicity plot of CatSper (Fig. 1b) predicted that it contained six-transmembrane-spanning domains. CatSper contains a remarkable abundance of histidine residues in its amino terminus (49/250 amino acids), which might be involved in the well-known pH regulation of sperm motility⁷. Human CatSper exhibits a high degree of homology (55% identity/72% similarity) with its mouse counterpart, especially in the transmembrane domains and the histidine-rich region. The transmembrane domains share 81% identity/93% similarity, and the pore region shares 89% identity/100% similarity. Low stringency screening of a mouse testis cDNA library with human CatSper detected no genes of higher similarity.

CatSper is localized to the principal piece

CatSper messenger RNA was detected in testis as a single band of roughly 2.6 kilobases. The size of the mRNA was similar to that of the cDNA clones that we isolated, suggesting that the gene product was a full-length transcript and not a truncated form of a much larger conventional Ca_v channel. CatSper mRNA was not detected in any of 8 mouse or 16 human organs, including brain, heart, spleen, lung and skeletal muscle (Fig. 2a). Furthermore, a CatSper mRNA probe recognized only testis in a dot blot of 50 human tissue mRNAs (Fig. 2b).

A polyclonal antibody directed against the first 150 amino acids of the amino terminus recognized CatSper protein (Fig. 2c). The antibody specifically labelled tetracycline-induced CatSper proteins from HEK-293 cells and native CatSper in testis, sperm (Fig. 2c) and spermatocytes. Consistent with the results of the northern blots, the antibody did not recognize similar size proteins from brain, heart or kidney (data not shown).

We used indirect immunofluorescence to determine the subcellular localization of sperm CatSper. The protein was highly localized to the principal piece of the sperm tail (Fig. 3a). CatSper was presumably localized to the plasma membrane because the principal piece does not contain intracellular organelles (such as endoplasmic reticulum). To determine the precise location of the protein in the sperm tail, we performed indirect immunogold electron microscopy. CatSper immunoreactive gold particles were detected in the plasma membrane above the fibrous sheath in the sperm principal piece (Fig. 3b). Consistent with the immunofluorescence staining, few immunoreactive gold particles were detected in either the head or midpiece. The detection of gold particles on the cytoplasmic face of the plasma membrane (Fig. 3b, middle) supports the predicted intracellular localization of the N terminus of CatSper.

Targeted disruption of CatSper

We disrupted CatSper in embryonic stem cells by homologous recombination to study its function *in vivo* (Fig. 4a). The second exon encoding the first putative transmembrane domain was replaced with an IRES–LacZ sequence followed by the neomycin resistance gene.

To generate chimaeras, embryonic stem cells carrying a mutant copy of the gene were injected into blastocysts and implanted in pseudo-pregnant mice. Mutant mice were generated from the mating of offspring. Disruption of the gene was confirmed by polymerase chain reaction (PCR) (Fig. 4b), western blotting (Fig. 4c), immunostaining (Fig. 4d) and immunogold electron microscopy (data not shown).

CatSper is required for male fertility

The genotypes of heterozygous (+/–) male and female offspring

exhibited roughly mendelian proportions (240 heterozygous, 106 wild-type and 110 mutant), suggesting that the mutation did not affect embryonic development. CatSper^{–/–} (mutant) mice were indistinguishable from their wild-type littermates in survival rates, appearance and gross behaviour.

Homozygous (–/–) females mated with heterozygous (+/–) or wild-type (+/+) males were fertile. Homozygous male mutants mated with wild-type females displayed mounting behaviour indistinguishable from that of wild-type males. The similar frequencies of vaginal plugs noted in mated females supported the normal mating behaviour of wild-type and mutant mice. However, the mutant males engendered no pregnancies over a period longer than 2 months (up to 9 months, $n = 26$ females, $n = 13$ males). In contrast, the wild-type littermate males were 100% fertile ($n = 8$ females, $n = 4$ males; Fig. 5a).

Body and testis weights of the mutant mice were not different from those of their wild-type counterparts (Fig. 5b). Sperm counts from mutant and wild-type caudal epididymis were not significantly different, and the sperm were cytologically indistinguishable (Fig. 5c). The absence of spermatogenesis defects in CatSper^{–/–} mice was supported by the lack of morphological differences between wild-type and mutant mouse testes (Fig. 5d).

A major effort was made to determine the precise electrophysiological characteristics of the wild-type and mutant sperm. As

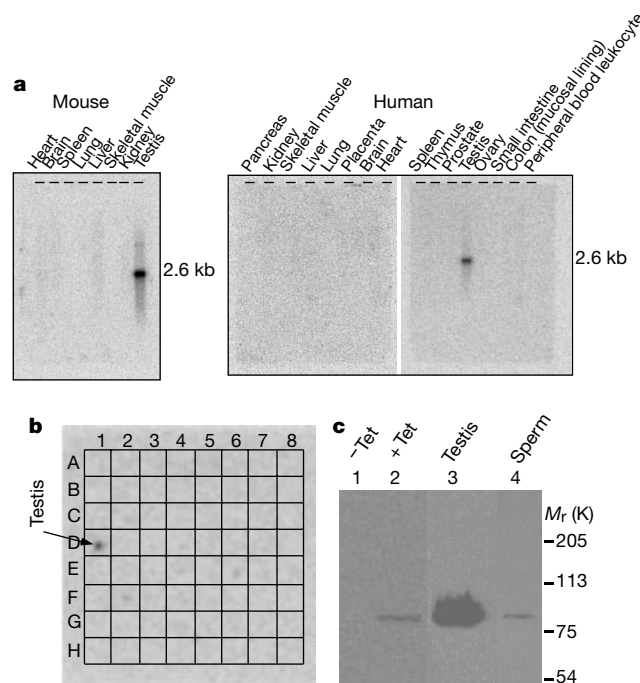


Figure 2 The restricted expression pattern of CatSper in testis. **a**, Northern blot from mice (left) and human (right) organs. **b**, Dot blot of mRNAs from 50 human tissues (grid superimposed). The tissues in the blot are from left to right: row A, whole brain, amygdala, caudate nucleus, cerebellum, cerebral cortex, frontal lobe, hippocampus, medulla oblongata; row B, occipital lobe, putamen, substantia nigra, temporal lobe, thalamus, nucleus accumbens, spinal cord; row C, heart, aorta, skeletal muscle, colon, bladder, uterus, prostate, stomach; row D, testis, ovary, pancreas, pituitary gland, adrenal gland, thyroid gland, salivary gland, mammary gland; row E, kidney, liver, small intestine, spleen, thymus, peripheral leukocyte, lymph node, bone marrow; row F, appendix, lung, trachea, placenta; row G, fetal brain, fetal heart, fetal kidney, fetal liver, fetal spleen, fetal thymus, fetal lung; and row H, negative controls. **c**, Western blot of CatSper protein. Lanes 1 and 2 were loaded with total protein from HEK-293 cells stably expressing inducible CatSper without (–Tet) and with (+Tet) tetracycline induction. Lanes 3 and 4 were loaded with testis membrane protein and sperm total protein, respectively. M_r , relative molecular mass.

detailed below, hundreds of attempts to measure whole-cell currents of both wild-type and mutant sperm under a variety of conditions and configurations were unsuccessful, consistent with the experience of other investigators. But we could record voltage-gated Ca^{2+} -channel currents from mutant and wild-type spermatocytes. The measured Ca^{2+} currents were consistent with previous data^{12,13}, and no significant differences between wild-type and mutant mouse spermatocyte inward currents were detected (Fig. 5e). No inward currents were induced by cyclic nucleotides in wild-type or mutant spermatocytes (data not shown). Thus, CatSper is not a main component of the spermatocyte voltage-sensitive Ca^{2+} currents and is not required for the development of sperm from spermatocytes.

CatSper is required for normal sperm motility

A closer examination of live sperm under light microscopy revealed a significant difference between the mutant and wild-type sperm. Sperm from wild-type mice displayed vigorous beating in the tail region and progressive directed movement. By contrast, CatSper^{-/-} sperm were sluggish and displayed less directed movements.

Most notably, mutant sperm lacked the vigorous beating and bending in the tail region. Computer-assisted sperm analysis showed that the mutant sperm's main motility parameters of path velocity, progressive velocity and track speed were impaired significantly (Fig. 6a). Thus, disrupting the CatSper gene results in markedly reduced sperm motility.

CatSper is required to penetrate the egg

In vitro fertilization (IVF) assays were performed to test CatSper^{-/-} sperm's ability to fertilize eggs. Superovulated wild-type female mature eggs incubated with capacitated wild-type or mutant sperm were examined after 24 h. The appearance of two-cell-stage embryos was taken to indicate successful penetration of sperm and

subsequent activation of fertilization. Whereas 81% of eggs (66/81) were fertilized by wild-type sperm, no eggs (0/79) were fertilized by CatSper^{-/-} sperm (Fig. 6b). The eggs fertilized by wild-type sperm developed into two-cell-stage embryos, but those incubated with mutant sperm remained at the single-cell stage (Fig. 6c). Some CatSper^{-/-} sperm adhered to the eggs but did not seem to be able to penetrate the eggs, possibly owing to impaired motility.

During the course of natural fertilization, sperm penetrate the surrounding layer of zona pellucida proteins and fuse to the egg's plasma membrane. To examine whether the mutant sperm retained the ability to fuse with the plasma membrane and activate fertilization, we incubated wild-type and mutant sperm with eggs whose outer layers had been enzymatically removed (zona-pellucida-free eggs). Under such conditions, sperm from both CatSper^{+/+} and CatSper^{-/-} mice were capable of fertilizing the eggs (Fig. 6b, c). Thus, we conclude that CatSper is required for the sperm to be able to penetrate the egg outer layers, but not for egg activation.

CatSper is required for cAMP-induced Ca^{2+} influx

Cyclic nucleotides have been widely implicated in sperm motility^{6-8,17}. To elucidate the function of CatSper, we expressed mouse and human CatSper in various heterologous expression systems including *Xenopus* oocytes, and HEK-293 and CHO-K1 cells. In over 250 patch-clamp studies, we failed to detect a significant current resulting from CatSper expression, alone or in combination with expressed CNG β -channel subunits (CNG4 or CNG6). Although CatSper protein was detected in the expression system (Fig. 2c), no currents could be elicited by changes in voltage, pH, osmolarity and/or cyclic nucleotide concentration. As a positive control, we measured cyclic-nucleotide-activated channel currents after expression of the CNG α -subunit (CNG1) and β -subunit (CNG6).

We also attempted to compare the currents of wild-type and

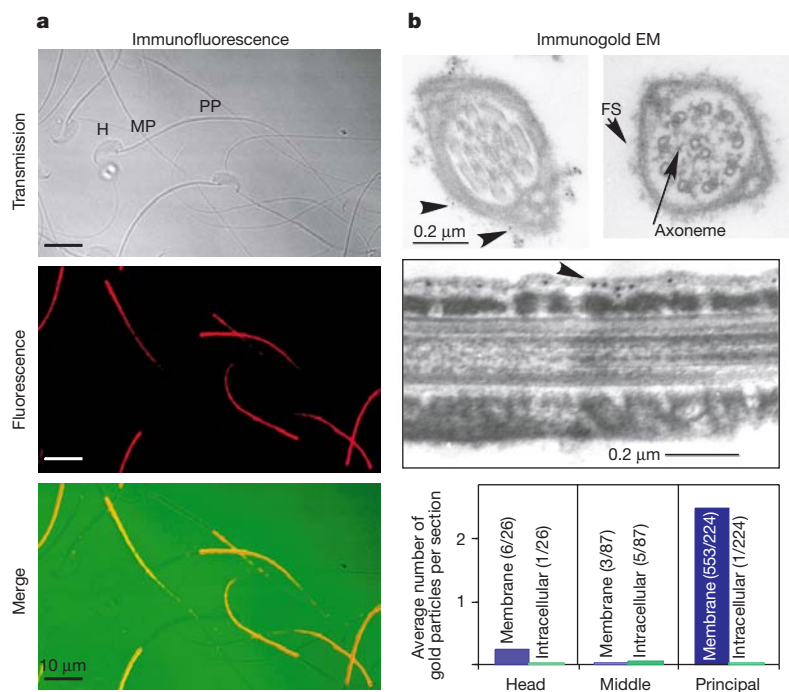


Figure 3 Localization of CatSper to the plasma membrane of the principal piece of sperm. **a**, Immunostaining of mouse sperm. Top, phase contrast. The head (H), midpiece (MP) and principal piece (PP) regions of the sperm are indicated. Middle, immunofluorescence. Bottom, merged signal where immunofluorescence is labelled red and transmission green. **b**, Sperm immunogold-labelled electron microscopy. Top left, cross-section through the principal piece; arrows indicate gold particles. Top right, fibrous sheath (FS).

Middle, longitudinal section. The immunoreactive gold particles are at the cell membrane cytoplasmic face. Bottom, distribution of gold particles along the sperm head, midpiece and principal piece. Gold particles located outside the fibrous sheath were counted as being membrane localized, and inside the fibrous sheath as intracellular. The numbers of sections counted and particles detected are indicated in parentheses.

mutant sperm head and tail membranes. As reported by others¹⁸, the success rate for obtaining stable patches from sperm was very low, presumably owing to the small diameter (0.5 μm) and the geometry of sperm tails. Out of about 900 attempts under various conditions, less than 30 gigaseals were formed and none was a stable whole-cell recording. Although no differences between wild-type and mutant cell-attached patches were noted, the low success rate prevented any reliable analysis. We therefore chose to examine sperm tail intracellular Ca^{2+} dynamics, which can be readily monitored with Ca^{2+} indicators^{11,19}.

We loaded wild-type and mutant sperm with Fluo-4, and measured their fluorescence by single-photon confocal microscopy. Application of cell-membrane-permeant cAMP (1 mM 8-Br-cAMP) to the bath solution containing caudal epididymal wild-

type sperm initiated a rapid rise in sperm head and tail intracellular calcium concentration ($[\text{Ca}^{2+}]_i$) (Figs 7a, b, and 8). This $[\text{Ca}^{2+}]_i$ signal was not detected in Ca^{2+} -free bath medium (data not shown), suggesting that the Ca^{2+} influx was mediated by a plasma membrane channel. Membrane-permeant cGMP (1 mM 8-Br-cGMP) also initiated Ca^{2+} influx in sperm. The cAMP- and cGMP-induced $[\text{Ca}^{2+}]_i$ increase was not blocked by incubating sperm with the nonspecific kinase inhibitor staurosporine (1 μM , 10 min), suggesting that cyclic-nucleotide-dependent protein kinases are not essential for channel activation.

Adding sodium bicarbonate (20 mM), a physiologically relevant stimulus to sperm⁸, to the bath induced an increase in $[\text{Ca}^{2+}]_i$ that was smaller than that observed with cAMP or cGMP (data not shown). A significant cyclic-nucleotide-induced $[\text{Ca}^{2+}]_i$ rise was also detected in testis-derived sperm but not in spermatids (nor in developing elongated spermatids with short tails; data not shown), suggesting that the Ca^{2+} response to cAMP is a property of the fully differentiated sperm. In $\text{CatSper}^{-/-}$ sperm, neither cAMP nor cGMP elicited a significant Ca^{2+} influx (Figs 7d, e, and 8). As a control for sufficient dye loading, progesterone initiated a large Ca^{2+} influx in both mutant (Fig. 7f) and wild-type sperm. We conclude that CatSper is required for the cAMP- and cGMP-induced Ca^{2+} influx in sperm.

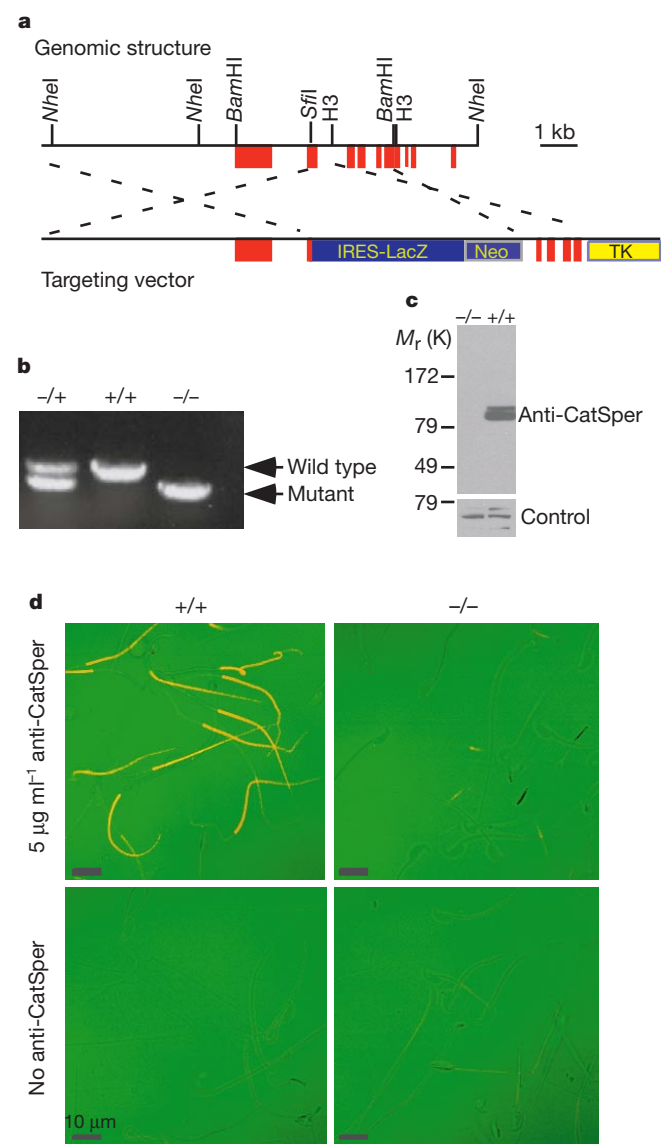


Figure 4 Targeted disruption of *CatSper*. **a**, Partial genomic structure of mouse *CatSper* and targeting vector. Filled red boxes, exons; thin lines, introns. **b**, PCR genotyping confirmed gene disruption. Tail genomic DNA was amplified with primers specific for *CatSper*^{+/+} and *CatSper*^{-/-}. **c**, *CatSper* protein was absent in mutant mice by western blot. Roughly equal amounts of membrane protein from wild-type and mutant testes were blotted with anti-*CatSper* antibody. An unrelated antibody controlled for equal protein loading. **d**, Immunostaining of wild-type (+/+) and mutant (-/-) sperm. Primary antibody was omitted in the control. The transmission signal is green, immunofluorescence is red, and overlap is orange.

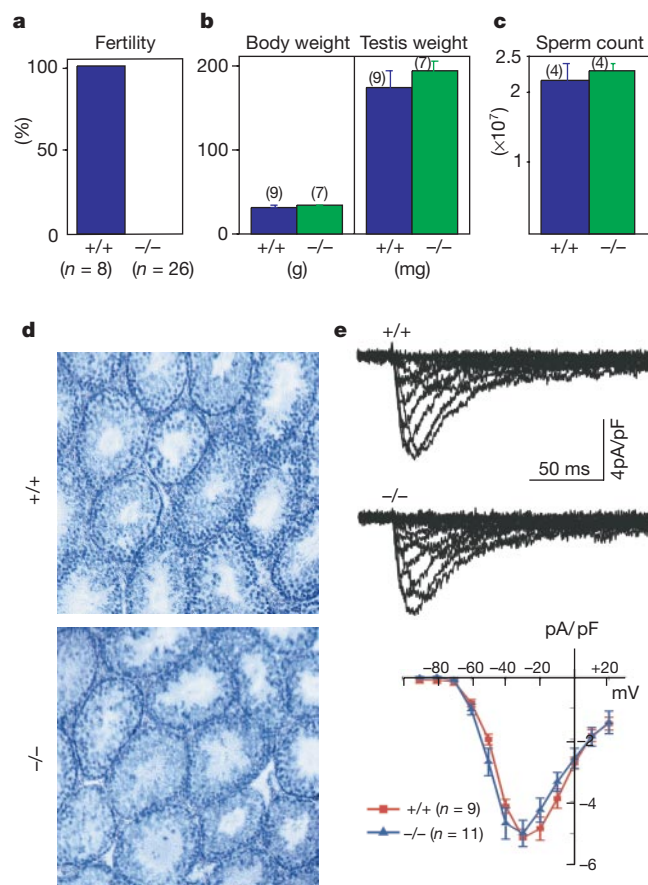


Figure 5 Male infertility caused by *CatSper* disruption. **a**, Fertility of *CatSper*^{+/+} and *CatSper*^{-/-} males. **b–d**, Mice body and testis weight (**b**), sperm count (**c**), and testis histology (**d**) were not significantly different. Testes sections were stained with haematoxylin/eosin. **e**, Ca^{2+} -channel currents were indistinguishable between wild-type (+/+) and mutant (-/-) spermatocytes in range of activation, amplitude and kinetics. Currents were elicited by test pulses from -90 to +20 mV ($\Delta 10$ mV). Bottom, averaged I/V relationships of inward currents. The holding potential was -100 mV; the average cell capacitances were 10.8 ± 1.4 pF (+/+, n=9) and 12.9 ± 0.9 pF (-/-, n=11).

Discussion

We have cloned a six-transmembrane-spanning putative cation channel, CatSper, that is restricted predominantly or completely to the principal piece of the sperm tail. This putative channel is required for male fertility in mice. CatSper mRNA and protein is present only in testis; immunocytochemistry localized CatSper specifically to the principal piece of mature sperm, and immunogold labelling showed that it is associated primarily with the plasma membrane. The complete lack of intracellular organelles in the principal piece suggests that CatSper cannot be an intracellular channel.

CatSper^{-/-} sperm are poorly motile and unable to fertilize eggs with an intact zona pellucida, but can fertilize eggs in which this membrane has been removed. Finally, on the basis of confocal microscopic measures of Ca²⁺ influx, CatSper is required for cyclic-nucleotide-mediated Ca²⁺ entry in the sperm tail. Taking these observations together, we propose that CatSper is a unique cation channel protein required for normal sperm motility and, in particular, for sperm penetration of the zona pellucida.

Although CatSper is expressed in heterologous systems, as revealed by the presence of protein, we were unable to show that the putative channel formed a functional channel in these non-sperm cells. Co-expression of CatSper and the CNG β -subunit, with or without cyclic nucleotide second messengers, did not result in measurable currents. Voltage-gated Ca²⁺ currents appeared normal in CatSper^{+/+} and CatSper^{-/-} spermatocytes, and CatSper protein did not contribute measurable inward or outward currents stimulated by voltage, pH changes, cyclic nucleotides or osmotic changes. In mature sperm, CatSper protein was localized to the principal

piece of the sperm tail—an area that is inaccessible to current patch-clamp methodologies. However, Ca²⁺ imaging of mature sperm from normal and mutant mice revealed that the protein is required for cyclic-nucleotide-stimulated Ca²⁺ influx. The simplest interpretation of the data is that CatSper protein comprises all or part of a tetrameric Ca²⁺-permeant channel that is gated directly or indirectly by cyclic nucleotides.

The failure to express the channels in heterologous systems suggests that accessory proteins in the principal piece of sperm are not present in any of the three expression systems we used. Alternatively, the proteins may be present in heterologous cells or spermatocytes, but might require structural attributes of the principal piece for their functional organization. Given the low abundance of principal piece tissue, direct biochemical approaches to isolate the putative accessory subunit are not currently feasible. Yeast two-hybrid approaches to find accessory proteins have also been unsuccessful. In a preliminary approach to finding potential accessory subunits, sperm-associated cyclic-nucleotide-gated channels were co-expressed with CatSper but did not yield measurable currents. Clearly, innovative approaches will be required to find the constituents of the presumed functional channel complex.

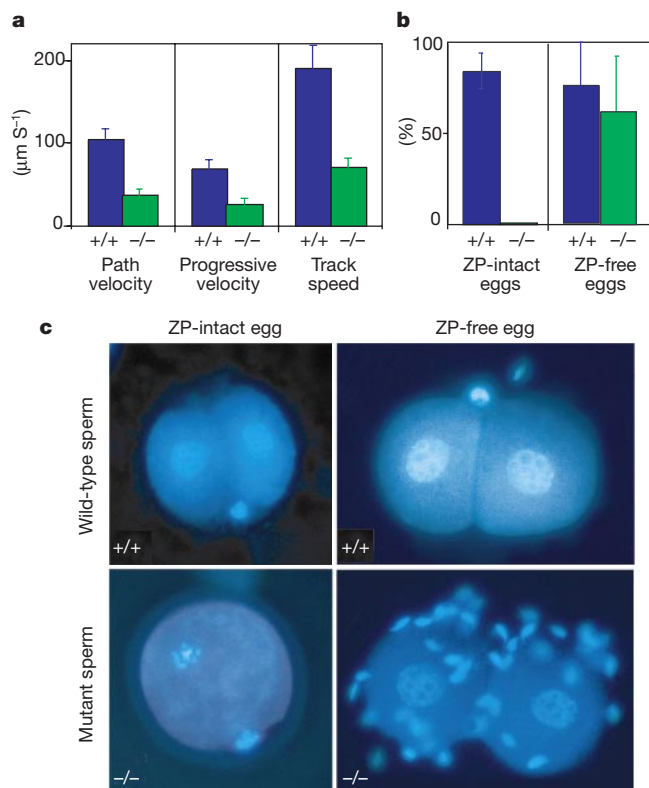


Figure 6 CatSper^{-/-} sperm motility and *in vitro* fertilization defects. **a**, Quantified sperm motility: path velocity, progressive velocity and track speed. **b**, *In vitro* fertilization (IVF) rates of mutant and the wild-type sperm with zona pellucida (ZP)-intact and ZP-free eggs (three pairs of mutant and wild-type mice each). **c**, Eggs with and without the ZP were incubated with CatSper^{+/+} or CatSper^{-/-} sperm. After 24 h, cells were stained with Hoechst 33342 for chromatin analysis. Note the decondensed chromatin in the two nuclei of the individual blastomeres at the two-cell stage.

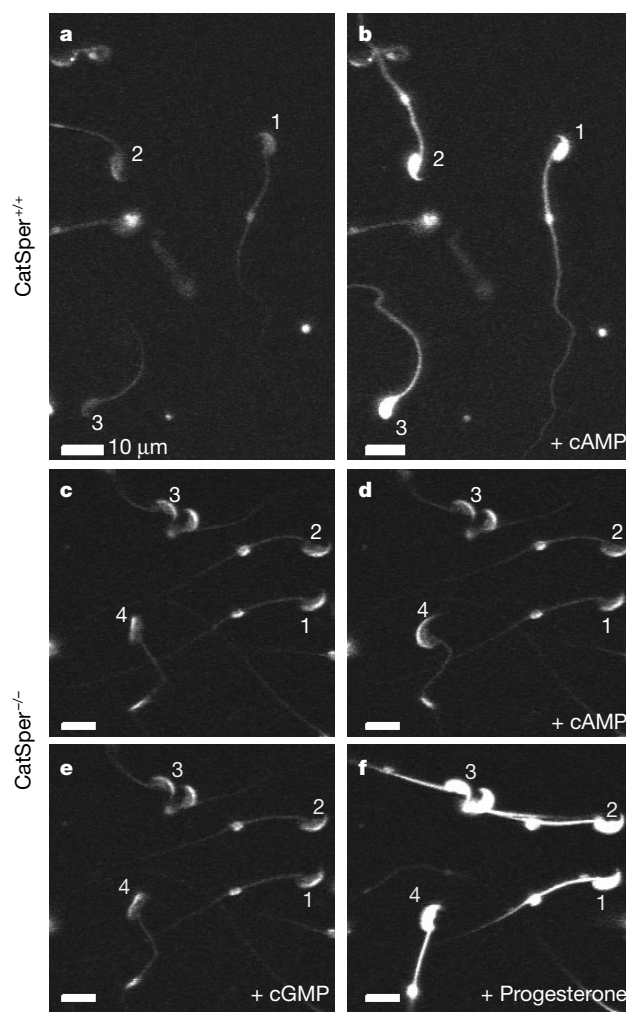


Figure 7 cAMP-induced calcium influx defect in CatSper^{-/-} sperm. **a**, **b**, Fluo-4 fluorescence images of wild-type sperm before (**a**) and after (**b**) addition of 1 mM cell-permeant cAMP (8-Br-cAMP). **c**–**f**, Incubation in cell-permeant cAMP (1 mM) or cGMP (1 mM) for 10 s did not induce Ca²⁺ influx in mutant (CatSper^{-/-}) sperm. Progesterone (~1 mM), which was used as a positive control for Ca²⁺ responses, initiated a large Ca²⁺ increase in both mutant (**f**) and control (not shown) sperm. Numbers identify individual spermatozoa.

CatSper represents a unique class of putative ion channel proteins. With the pore features of a Ca^{2+} -selective ion channel, a putative voltage-sensitive S4 domain, and the single repeat structure of K_v , CNG, HCN and TRP channels, CatSper may represent an evolutionary transition channel before gene duplication gave rise to the four-repeat structure of Ca_v channels. The activation mechanism for CatSper is solely a matter of speculation. Cyclic nucleotides regulate sperm motility and are themselves dynamically regulated⁶; however, there are no putative cyclic-nucleotide-binding regions in the primary structure of CatSper. If CatSper is not bound directly by nucleotides, then unknown sperm principal-piece-specific accessory subunits might mediate its regulation.

The activation of CatSper mediates a rise in intracellular Ca^{2+} levels in the sperm tail. The rise of Ca^{2+} regulates sperm motility,

presumably through the motor proteins of the tail. If Ca^{2+} -sensitive adenylyl cyclases are present in the sperm, Ca^{2+} entry could accelerate the rise of cAMP^{2,6,8}, thus providing a positive feedback loop that would amplify cyclic-nucleotide signalling. Such a positive feedback loop may be important in the most vigorous sperm tail beating required for the sperm to penetrate the egg outer layers. Whatever the mechanism of activation, the activity of CatSper seems to depend on the specific structure of sperm tails, and the principal piece in particular. Such a hypothesis explains our finding that cyclic nucleotides induce significant Ca^{2+} influx in testis sperm, but not in spermatids that lack a significant tail.

The human CatSper gene is a potential target for male infertility screening and treatment. Human CatSper is also an ideal target for potential contraceptive drugs. As the *in vitro* fertilization results suggest that CatSper^{-/-} sperm cannot fertilize eggs, a specific blocker of CatSper might be effective when taken by either men or women. The restricted localization of CatSper to mature sperm means that a specific blocker should not affect other tissues, and thus side-effects should be low or nonexistent. The normal development and behaviour (including sexual) of the mutant mice supports such a prediction. Finally, as the channel represents a new structure, it may be an excellent target for channel agonists or antagonists. □

Methods

Cloning of CatSper

High-voltage-gated Ca^{2+} -channel protein sequences were used to search the EST database. One EST (AA416682) sequence was analysed and found to be similar to voltage-gated Ca^{2+} channels. We used primers derived from this sequence in PCR after reverse transcription of RNA (RT-PCR) to obtain a cDNA fragment. The fragment was used as a probe to localize the main source of expression in testis. We used cDNA library screening and rapid amplification of cDNA ends to clone the cDNAs containing the whole open reading frame (ORF) from human and mouse testes. The start of the ORF was confirmed by an in-frame stop codon upstream of the first ATG. cDNA constructs for expression in mammalian cells were made by PCR using Pfu polymerase. The constructs were sequenced completely and contained no mutations. Northern and dot blots were from Clontech.

Protein methods

Affinity-purified polyclonal antibody was generated against a glutathione S-transferase (GST) fusion protein of the N-terminal 150 amino acids. Western blots were carried out with $2 \mu\text{g ml}^{-1}$ anti-CatSper antibody. Immunostaining with $5 \mu\text{g ml}^{-1}$ anti-CatSper antibody was detected with a rhodamine-conjugated anti-rabbit secondary antibody. Anti-CatSper ($10 \mu\text{g ml}^{-1}$) and 10-nm gold-conjugated protein A particles were used in immuno-electron microscopy; 60-nm sections were cut from embedded cells. We used CatSper^{-/-} sperm as controls in all the experiments to confirm the specificity of the antibodies.

Generation of knockout mice

Genomic bacterial artificial chromosomes clones containing the CatSper gene were isolated and analysed by restriction digestion and sequencing. A targeting vector (Fig. 4a) was constructed to contain 5 kb (left arm), IRES-LacZ followed by a Neo-resistance gene (middle), and 1.7 kb followed by TK as a negative selection marker (right arm). Embryonic stem cells derived from 129/SvJ mice were transfected with linearized DNA, selected and analysed for correct targeting. We used correctly targeted embryonic stem cells to generate chimaeric mice, and crossed these mice with C57BL/6J to obtain heterozygous mutants. Mice used in the study were the offspring of crosses between F₁ and/or F₂ generations (129/SvJ/C57BL/6J genetic background). Wild-type littermates were used as controls in the IVF experiments.

In vitro fertilization

Sperm were collected from cauda epididymis and capacitated *in vitro* for 2 h. Eggs were collected from mature wild-type females (~6-weeks old) synchronized with 10 units of PMSG (pregnant mare serum gonadotropin) and 10 units of hCG (human chorionic gonadotropin) 46–48 h and 14 h before collection, respectively. We carried out IVF using standard protocols²⁰. Eggs were incubated with roughly 10^5 CatSper^{+/+} or 10^5 CatSper^{-/-} sperm for 4 h at 37 °C, and unbound sperm were washed away. After a 20-h incubation at 37 °C, eggs were examined for the presence of the two-cell-stage embryos as an indication of successful fertilization. Some embryos were fixed and stained with Hoechst 33342 for imaging immediately after IVF. The zona pellucida was removed by treatment with hyaluronidase (80 IU ml^{-1}) followed by perfusion in Tyrode's solution for 30–40 s.

Electrophysiology

Spermatocyte preparation and whole-cell current recording were performed essentially as described^{12,21}. The pipette solution contained (in mM): 120 Cs⁺, 60 glutamic acid,

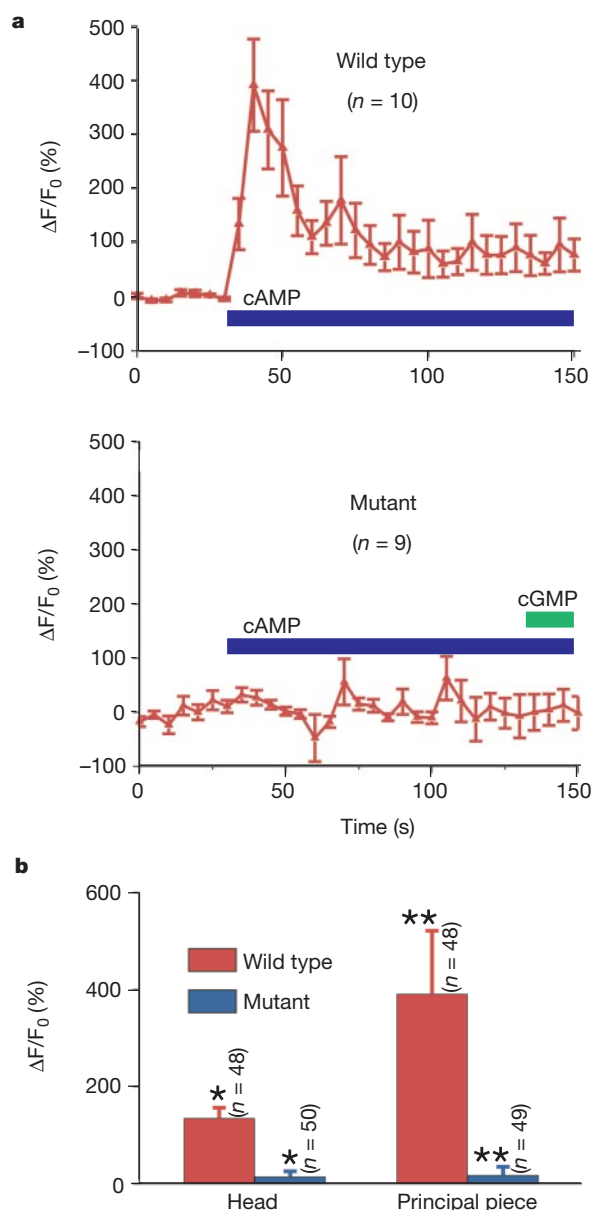


Figure 8 Summary of cyclic-nucleotide-induced Ca^{2+} changes. **a**, Time courses of cAMP-induced Ca^{2+} changes in the principal pieces of wild-type (CatSper^{+/+}) and mutant (CatSper^{-/-}) sperm. **b**, Mean fluorescence change in sperm heads and principal pieces from CatSper^{+/+} and CatSper^{-/-} mice. Fluorescence was normalized to the average of the first seven sampling points of the pre-stimulation trace. ** $P < 0.005$; * $P < 0.001$; indicating statistical significance.

20 TEA-Cl, 5 MgCl₂, 3 Mg-ATP, 10 EGTA, 10 HEPES and 5 D-glucose (pH 7.4). The bath contained (in mM): 135 NaCl, 5 KCl, 10 CaCl₂, 10 sodium lactate, 10 sodium pyruvate, 10 glucose and 30 HEPES (pH 7.4). We subtracted leak currents using an online P/4 protocol. Data were filtered at 1 kHz and corrected for junction potentials.

Ca²⁺ imaging

Sperm were prepared in HS medium¹⁴ and loaded with 10 μM Fluo-4-AM and 0.05% Pluronic F-127 for 30 min at room temperature. Washed cells were seeded onto Cell-Tak coated coverslips. Attached motile sperm were imaged in nonconfocal mode to avoid motion-induced fluorescence change artefacts. We used the 488-nm wavelength beam of an argon–krypton laser for excitation, and monitored fluorescence emission at 512 nm (Zeiss LSM 410).

Received 5 July; accepted 16 August 2001.

1. Wassarman, P. M., Jovine, L. & Litscheer, E. S. A profile of fertilization in mammals. *Nature Cell Biol.* **3**, E59–E64 (2001).
2. Yanagimachi, R. in *The Physiology of Reproduction* (eds Knobil, E. & Neill, J. D.) 189–315 (Raven, New York, 1994).
3. O'Toole, C. M., Arnoult, C., Darszon, A., Steinhardt, R. A. & Florman, H. M. Ca²⁺ entry through store-operated channels in mouse sperm is initiated by egg ZP3 and drives the acrosome reaction. *Mol. Biol. Cell* **11**, 1571–1584 (2000).
4. Jungnickel, M. K., Marrero, H., Birnbaumer, L., Lemos, J. R. & Florman, H. M. Trp2 regulates entry of Ca²⁺ into mouse sperm triggered by egg ZP3. *Nature Cell Biol.* **3**, 499–502 (2001).
5. Bedford, J. M. Mammalian fertilization misread? Sperm penetration of the eutherian zona pellucida is unlikely to be a lytic event. *Biol. Reprod.* **59**, 1275–1287 (1998).
6. Tash, J. S. in *Controls of Sperm Motility: Biological and Clinical Aspects* (ed. Gagnon, C.) 229–240 (CRC, Boca Raton, 1990).
7. Darszon, A., Labarca, P., Nishigaki, T. & Espinosa, F. Ion channels in sperm physiology. *Physiol. Rev.* **79**, 481–510 (1999).
8. Hyne, R. V. & Garbers, D. L. Calcium-dependent increase in adenosine 3',5'-monophosphate and induction of the acrosome reaction in guinea pig spermatozoa. *Proc. Natl Acad. Sci. USA* **76**, 5699–5703 (1979).
9. Serrano, C. J., Trevino, C. L., Felix, R. & Darszon, A. Voltage-dependent Ca²⁺ channel subunit expression and immunolocalization in mouse spermatogenic cells and sperm. *FEBS Lett.* **462**, 171–176 (1999).
10. Weyand, I. *et al.* Cloning and functional expression of a cyclic-nucleotide-gated channel from mammalian sperm. *Nature* **368**, 859–863 (1994).

11. Wiesner, B. *et al.* Cyclic nucleotide-gated channels on the flagellum control Ca²⁺ entry into sperm. *J. Cell Biol.* **142**, 473–484 (1998).
12. Santi, C. M., Darszon, A. & Hernandez-Cruz, A. A dihydropyridine-sensitive T-type Ca²⁺ current is the main Ca²⁺ current carrier in mouse primary spermatocytes. *Am. J. Physiol.* **271**, C1583–C1593 (1996).
13. Arnoult, C., Cardullo, R. A., Lemos, J. R. & Florman, H. M. Activation of mouse sperm T-type Ca²⁺ channels by adhesion to the egg zona pellucida. *Proc. Natl Acad. Sci. USA* **93**, 13004–13009 (1996).
14. Wennemuth, G., Westenbroek, R. E., Xu, T., Hille, B. & Babcock, D. F. Cav2.2 and Cav2.3 (N- and R-type) Ca²⁺ channels in depolarization-evoked entry of Ca²⁺ into mouse sperm. *J. Biol. Chem.* **275**, 21210–21217 (2000).
15. Ertel, E. A. *et al.* Nomenclature of voltage-gated calcium channels. *Neuron* **25**, 533–535 (2000).
16. Ellinor, P. T., Yang, J., Sather, W. A., Zhang, J. F. & Tsien, R. W. Ca²⁺ channel selectivity at a single locus for high-affinity Ca²⁺ interactions. *Neuron* **15**, 1121–1132 (1995).
17. Aoki, F., Sakai, S. & Kohmoto, K. Regulation of flagellar bending by cAMP and Ca²⁺ in hamster sperm. *Mol. Reprod. Dev.* **53**, 77–83 (1999).
18. Espinosa, F. *et al.* Mouse sperm patch-clamp recordings reveal single Cl[−] channels sensitive to niflumic acid, a blocker of the sperm acrosome reaction. *FEBS Lett.* **426**, 47–51 (1998).
19. Kobori, H., Miyazaki, S. & Kuwabara, Y. Characterization of intracellular Ca²⁺ increase in response to progesterone and cyclic nucleotides in mouse spermatozoa. *Biol. Reprod.* **63**, 113–120 (2000).
20. Perez, G. I. *et al.* Prolongation of ovarian lifespan into advanced chronological age by Bax-deficiency. *Nature Genet.* **21**, 200–203 (1999).
21. Arnoult, C., Lemos, J. R. & Florman, H. M. Voltage-dependent modulation of T-type calcium channels by protein tyrosine phosphorylation. *EMBO J.* **16**, 1593–1599 (1997).

Acknowledgements

We thank K. Wickman and W. Pu for advice on embryonic stem cell handling; G. Krapivinsky and X. Wei for assistance with antibody work and confocal microscopy, respectively; M. Ericsson for electron microscopy; H. Florman, D. Babcock, C. Santi and A. Darszon for advice on spermatocyte and sperm preparation; M. Biel and K.-W. Yau for CNG clones; and D. Garbers and T. Quill for helpful discussions and advice. B.N. was supported by Centro Internacional de Fisica, Lab. de Biofisica, Bogota, Columbia. This work was supported by the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to D.E.C. (e-mail: dclapham@rascal.med.harvard.edu). The GenBank accession numbers are AF407332 for mouse CatSper and AF407333 for human CatSper.

Tunable quantum tunnelling of magnetic domain walls

J. Brooke^{*†}, T. F. Rosenbaum^{*} & G. Aeppli[†]

^{*} The James Franck Institute and Department of Physics, The University of Chicago, Chicago, Illinois 60637, USA

[†] NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540, USA

Perhaps the most anticipated, yet experimentally elusive, macroscopic quantum phenomenon¹ is spin tunnelling in a ferromagnet², which may be formulated in terms of domain wall tunnelling^{3,4}. One approach to identifying such a process is to focus on mesoscopic systems where the number of domain walls is finite and the motion of a single wall has measurable consequences. Research of this type includes magnetotransport measurements on thin ferromagnetic wires⁵, and magnetization experiments on single particles^{6,7}, nanomagnet ensembles^{8–10} and rare-earth multilayers¹¹. A second method is to investigate macroscopic disordered ferromagnets^{12–15}, whose dynamics are dominated by domain wall motion, and search the associated relaxation-time distribution functions for the signature of quantum effects. But whereas the classical, thermal processes that operate in these experiments are easily regulated via temperature, the quantum processes have so far not been tunable, making difficult a definitive interpretation of the results in terms of tunnelling. Here we describe a disordered magnetic system for which it is possible to adjust the quantum tunnelling probabilities. For this material, we can model both the classical, thermally activated response at high temperatures and the athermal, tunnelling behaviour at low temperatures within a unified framework, where the domain wall is described as a particle with a fixed mass. We show that it is possible to tune the quantum tunnelling processes by adjusting the ‘mass’ of this particle with an external magnetic field.

Figure 1a depicts domain wall motion in the classical and quantum limits against the background of a fixed potential landscape, which pins the walls. In the classical case, indicated by the blue arrow, the domain wall moves over the potential barrier, thermally flipping spins as it advances. In the extreme quantum case (red arrow), the domain wall tunnels through the barrier, with the possibility in rare instances of flipping all barrier spins simultaneously. The problem illustrated in Fig. 1a reduces to the more familiar quantum barrier tunnelling problem (Fig. 1b) if the domain wall is associated with a particle of effective mass m , moving in a one-dimensional potential derived from the real three-dimensional pinning potential. The mass is a key parameter—heavy particles behave classically and remain more localized than light particles. Our experiments on the disordered ferromagnet $\text{LiHo}_{0.44}\text{Y}_{0.56}\text{F}_4$ show that Fig. 1b provides a detailed description of domain wall motion, with m a continuously tunable parameter.

LiHoF_4 is a tetragonal insulating ferromagnet, with a Curie temperature $T_C = 1.53$ K and an ordered moment along the crystal c axis. For magnetic fields H_t applied perpendicular to the c axis, the material becomes the experimental realization of the simplest quantum spin model, the Ising ferromagnet in a transverse field. The corresponding hamiltonian is

$$H = - \sum_{ij} J_{ij} \sigma_i^z \sigma_j^z - \Gamma_c \sum_i \sigma_i^x \quad (1)$$

where the σ s are Pauli spin matrices located at lattice sites i and j , the J_{ij} s are longitudinal couplings, and Γ_c is an effective transverse field, perpendicular to the Ising axis and proportional to H_t^2 for small H_t . In the $\Gamma_c = 0$ limit, atomically thin domain walls separate regions

with $\sigma_z = +1$ from $\sigma_z = -1$. This wall is dynamically stable for $\Gamma_c = 0$ because H commutes with σ_z ; the domain wall has infinite mass. For non-zero Γ_c , the commutator $[H, \sigma_z]$ no longer vanishes and wall motion can occur; the mass of the domain wall is now finite. Eventually, when Γ_c becomes comparable to J , there is a quantum critical point (occurring at $\Gamma_c = 5.3$ K for pure LiHoF_4)¹⁶ beyond which the Ising ferromagnetism is unstable. As $\Gamma_c \rightarrow \Gamma_c$, the domain walls broaden to fill the entire system (resulting in zero net moment), or equivalently, their mass approaches zero. In practice, even for zero applied Γ , the internal dipole fields of LiHoF_4 create an internal field Γ_i which slightly broadens the domain walls¹⁷ and cuts off the mass at large but finite value. The effective magnetic field is thus $\Gamma_e = \Gamma + \Gamma_i$. If suitable barriers to domain wall motion are introduced into LiHoF_4 , the domain wall dynamics can pass from classical (large m) to quantum mechanical (small m) limits within our measurement window via a simple increase of Γ from 0 towards Γ_c . We can insert such barriers by the random, partial substitution of non-magnetic Y for magnetic Ho, leading to quenched disorder that is ideal for pinning domain walls. With suitable magnetic dilution x , $\text{LiHo}_x\text{Y}_{1-x}\text{F}_4$ in a transverse field is a macroscopic system for which we can vary domain wall mass for a potential energy landscape dominated by fixed pinning centres.

We use both static and dynamic measurements to explore the nature of the ordered state. The technique of d.c. magnetometry reveals standard magnetization—longitudinal field (M – H) hysteresis loops characteristic of pinned domains in a ferromagnet, with widths of the order of tens of oersteds and a saturation magnetization $4\pi M \approx 200$ Oe (Fig. 2). Both raising temperature (thermal fluctuations) and increasing the transverse field (quantum fluctuations) serve to depin the domain walls and narrow the hysteresis loops.

The technique we use to probe domain wall motion is magnetic susceptibility, which measures the incremental changes in magnetization due to an infinitesimal oscillating field. For a ferromagnet, such changes correspond to the growth of domains polarized parallel to the field at the expense of domains with antiparallel polarization. The growth occurs via domain wall motion of the type

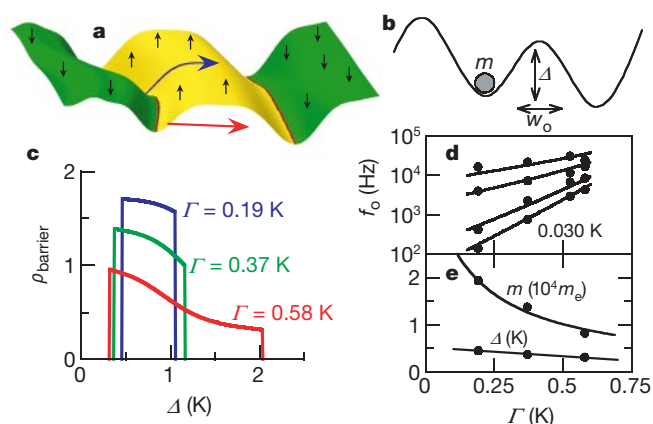


Figure 1 Domain wall tunnelling. **a**, Cartoon depicting motion of a domain wall separating regions of opposing spin orientation. The classical, thermally activated process is indicated with a blue arrow, and the quantum tunnelling route is shown in red. **b**, Sketch of the domain wall in **a** modelled as a particle in a one-dimensional potential with barrier height Δ and width w_0 . **c**, Experimentally obtained domain wall barrier distribution as a function of energy at a series of transverse fields, Γ . **d**, Frequency response of the weakest-pinned domain walls as a function of Γ at temperatures $T = 0.030, 0.070, 0.110$ and 0.150 K, in order of increasing response magnitude. Solid lines follow from equation (5). **e**, Best-fit values to equation (5) for the domain wall mass m and potential barrier height Δ of equation (2) for the weakest-pinned domain walls, as functions of Γ . The solid line through m is $\lambda(\Gamma + \Gamma_i)$, where $\lambda = (0.66 \pm 0.05) \times 10^4 m_e$ K and $\Gamma_i = 0.15 \pm 0.02$ K. To obtain real masses, the tunnelling distance, w_0 , has been set to the average spacing a between magnetic ions, 8.1 Å.

illustrated in Fig. 1a. If the motion is fast on the scale of the measuring frequency f , the susceptibility $\chi(f)$ will be frequency-independent, whereas if it is slow, the domains will not be able to follow the a.c. field and the in-phase part of χ will be reduced. The simplest motion is relaxational, and is characterized by a single relaxation time τ . For classical barrier hopping, τ will follow a thermally activated Arrhenius form, $1/\tau = f_m \exp(-\Delta/T)$, with microscopic attempt frequency f_m and barrier energy Δ . In the quantum limit, the rate at which the wall can tunnel through the same barrier will be the squared amplitude of the wavefunction on the other side of the barrier a distance w_0 away. For a square barrier, this is:

$$\frac{1}{\tau} = f_m e^{-2w_0 \sqrt{\frac{2m\Delta}{\hbar^2}}} \quad (2)$$

Thus, by measuring $\chi(f)$ as a function of T , we can monitor the evolution from classical Arrhenius behaviour to the T -independent quantum regime. For the random ferromagnet $\text{LiHo}_{0.44}\text{Y}_{0.56}\text{F}_4$, the domain wall dynamics will be defined not by a single τ , but by a distribution $\rho(\tau)$ of relaxation times, associated with a distribution of barriers Δ , as illustrated in Fig. 1c. Assuming that the different τ s are due to independent (non-interfering) processes, the response function becomes

$$\chi(f) = \int_0^\infty \rho(\tau) \chi_\tau(f) d\tau \quad (3)$$

where $\chi_\tau(f) = \chi_0/(1 - 2\pi i f \tau)$ is the Debye response¹⁸. Our strategy,

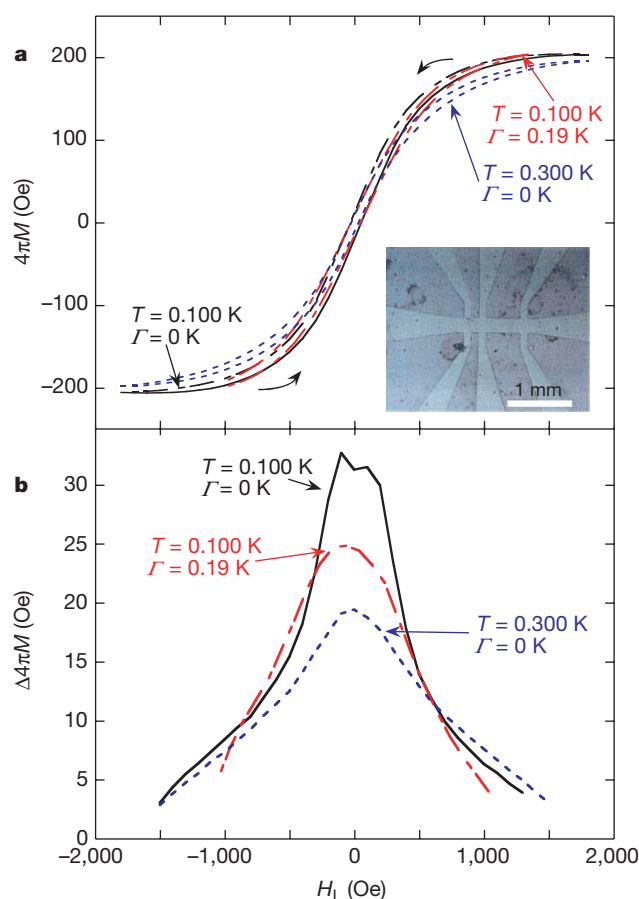


Figure 2 Static magnetization. **a**, Magnetization—longitudinal field (M - H) hysteresis loops after zero longitudinal field cooling at three temperatures and transverse fields, consistent with domain ordering: $T = 0.300$ K, $\Gamma = 0$ K (blue short dash); $T = 0.100$ K, $\Gamma = 0$ K (black solid line); $T = 0.100$ K, $\Gamma = 0.19$ K ($H_t = 5$ kOe) (red long dash). Inset, photograph of the InAs Hall bar assembly used for the magnetization measurements. Scale bar, 1 mm. **b**, The M - H loop widths of the curves in **a**.

therefore, is to measure $\chi(f)$, and then to examine the implications for the evolution of $\rho(\tau)$ with T and Γ .

We plot in Fig. 3a the phase diagram as a function of temperature and transverse field for single-crystal $\text{LiHo}_{0.44}\text{Y}_{0.56}\text{F}_4$. In the large T , small Γ classical limit, the system is a disordered ferromagnet with a Curie temperature, $T_C(\Gamma = 0) = 0.65$ K, while for $T = 0$ it has a quantum critical point at $\Gamma_C(T = 0) = 1.6$ K (ref. 19). The domain wall dynamics are encoded in the frequency-dependent magnetic susceptibility data of Fig. 3b. To analyse our results, we use equation (3) and choose $\rho(\tau)$ to be as simple as possible yet consistent with the observed data:

$$\rho(\tau) = \rho_0 \delta(\tau) + \rho_1 \tau^{-1} \Theta(\tau - \tau_1) \Theta(\tau_h - \tau) \quad (4)$$

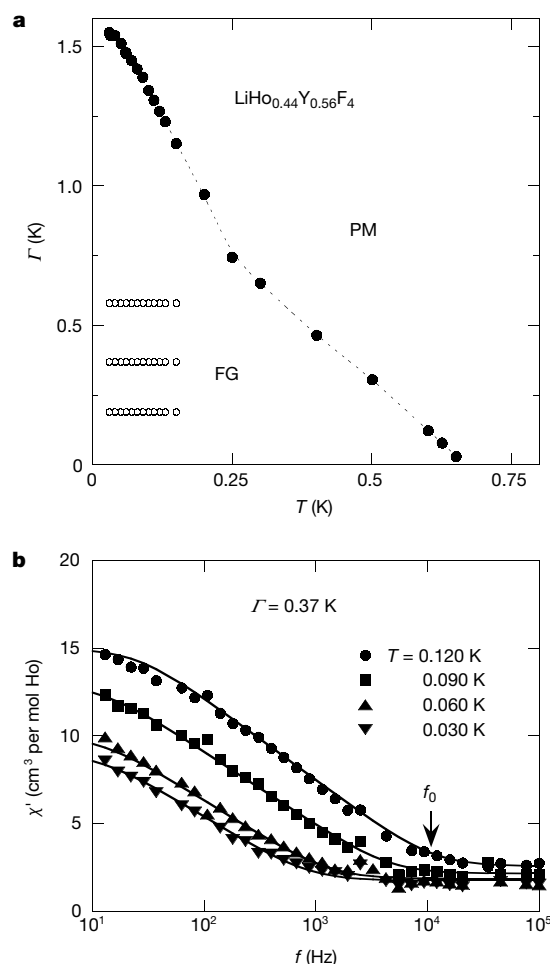


Figure 3 Phase diagram and spectral characterization of the ordered state. **a**, Transverse field—temperature (Γ - T) phase diagram for the disordered Ising magnet, with 44% of the sites occupied by (holmium) magnetic dipoles. The transverse field introduces tunnelling modes for magnetization that depress the temperature for spin freezing in a controllable fashion. Filled circles denote the phase boundary, measured by the cusp in susceptibility at 5 Hz, with the dashed line a guide to the eye. Open circles denote the values of Γ and T at which the measurement data in **b** and Fig. 4 were obtained. PM, paramagnet; FG, ferroglass. **b**, The spectroscopic response in the quantum limit is characterized by a logarithmic dependence of the real part of the magnetic susceptibility, χ' , over several decades below a characteristic frequency f_0 . An arrow highlights this crossover for $T = 0.120$ K. Reducing T decreases f_0 , as well as the amplitudes of both the logarithmic low- f and constant high- f terms. However, the effect of reducing T (that is, by similar increments δT) diminishes with decreasing T . Thus, the magnetic relaxation appears to approach a T -independent quantum limit on all measured frequency scales. Points are from measured data at $\Gamma = 0.58$ K, with error bars smaller than the symbol sizes, and lines are best-fit values to equations (3) and (4).

Here Θ is a Heaviside step function, prefactors ρ_0 and ρ_1 provide normalization, while the $1/\tau$ form of the second term yields the observed logarithmic divergence, with cut-offs at the low and high ends, τ_l and τ_h , dictated by the data (τ_h is only observable for a few curves at the highest T and Γ). The delta-function at very short times accommodates the flat spectral response at high frequencies.

Given the simplicity of $\rho(\tau)$, the parameters τ_l and τ_h summarize the entire measured dynamics of $\text{LiHo}_{0.44}\text{Y}_{0.56}\text{F}_4$, and consistently fitting the data to this form yields unique values for f_0 whose systematic behaviour as a function of Γ and T can then be studied. We focus on the crossover frequency $f_0 = 1/\tau_h$, corresponding to the fastest large-scale relaxation, and the most weakly pinned domain walls. Plotting f_0 against inverse T for several transverse fields yields Fig. 4, the central result of our experiment, which demonstrates explicitly the evolution from classical to quantum relaxation. At the higher temperatures, the domain wall relaxation follows the Arrhenius form with a universal (Γ - and T -independent) microscopic attempt frequency, $f_m = 2.2 \pm 0.2 \times 10^5$ Hz, and a Γ -dependent barrier energy Δ , shown in Fig. 1e. For increasing transverse field, the system approaches the disordered paramagnetic state, with the result that Δ is reduced linearly, $\Delta(\Gamma) = \Delta_0(1 - \Gamma/\Gamma_\Delta)$ with $\Delta_0 = 0.54 \pm 0.03$ K. The solid line through the data in Fig. 1e corresponds to this form, and extrapolates to zero at $\Gamma_\Delta = 1.3 \pm 0.2$ K, remarkably close to the quantum critical field $\Gamma_c(T = 0)$ beyond which the material is paramagnetic (Fig. 3a). Below $T \approx 0.1$ K, there are clear deviations in Fig. 4 from Arrhenius behaviour, with f_0 becoming temperature-independent as $T \rightarrow 0$. The crossover temperature to quantum behaviour increases with Γ ; at high transverse field, the increased tunnelling probability permits the quantum regime to extend to higher T . This observation, combined with the continued evolution down to the lowest temperatures of the Γ - T phase diagram of Fig. 3a, proves that the saturation in f_0 is intrinsic and not due to a thermal decoupling of the sample from the dilution refrigerator.

The transverse field is far more efficient at relaxing the system in the quantum regime than in the thermal regime. The comparative advantage is most apparent in Fig. 1d, where we show f_0 as a function of Γ at several values of T . At the base temperature of $T = 0.030$ K, tripling Γ from 0.19 K to 0.58 K increases f_0 by one-and-a-half orders of magnitude, while at $T = 0.150$ K it has a negligible effect. Having determined the barrier Δ at each Γ from the high-temperature regime, we are able to use the barrier tunnelling form equation (2) to extract the combination mw_0^2 from the measured f_0 . It is clear that mw_0^2 varies more rapidly with Γ than does Δ (Fig. 1e); the principal effect of the transverse field in this part of the Γ - T plane is to reduce the effective mass of the domain walls.

The barrier width w_0 , which characterizes the distances over which the domain walls must hop, should be fixed by the frozen impurity configuration and ought to be independent of Γ . Hence, we attribute the reduction of mw_0^2 with Γ to a reduction in m , describable by $m = \lambda(\Gamma + \Gamma_i)$, where λ is a constant (solid line in Fig. 1e).

The simplest expression for the entire T - and Γ -dependent relaxation is based on assuming that classical and quantum processes provide independent relaxation channels. Thus, we simply add the quantum and classical forms identified above, that is:

$$f = f_m \left\{ \exp(-\Delta_0(1 - \Gamma/\Gamma_\Delta)/T) + \exp\left(-A\sqrt{(\Gamma + \Gamma_i)^{-1}}\sqrt{\Delta_0(1 - \Gamma/\Gamma_\Delta)}\right) \right\} \quad (5)$$

Here f_m characterizes the response time of the most weakly-pinned domain walls and $A = 2w_0\sqrt{2\lambda/\hbar^2}$. This expression, although only first order in the most relevant parameters of the system and lacking cross terms, fits the data quite well in both T and Γ , as evidenced from the solid lines of Fig. 4. It is only valid deep in the ordered phase of Fig. 3a, with open circles denoting where measurements were taken.

A further test of the tunnelling and hopping particle model for domain wall motion (illustrated in Fig. 1b and encapsulated by equation (5)) is whether it applies to more than just the fastest processes, derived from the lowest barriers. In particular, can a single mass account for the entire spectral response in both the classical and quantum regimes exhibited for a single Γ in Fig. 3b? Indeed it can, as evidenced by the solid lines through the data. This three-parameter fit is derived from a fixed barrier distribution function $\rho(\Delta)$, shown in Fig. 1c, such that for each Δ and T , the relaxation time is given by equation (5). For the smallest $\Gamma = 0.19$ K, the distribution is narrow, with a width of the same order as the centroid. This sharpness at low Γ is consistent with ferromagnetic ordering, in spite of the significant (56%) disorder. Increasing Γ broadens the barrier distribution.

One important question concerns the number (N) of spins that are tunnelling coherently—have we seen macroscopic quantum tunnelling? The collective nature of the spin dynamics is borne out by the numbers. An individual (isolated or weakly coupled) spin would have a flipping rate of the order of $\Gamma/\hbar$, which for $\Gamma = 0.4$ K is 10^{10} Hz, five orders of magnitude larger than the measured f_0 . More directly, we can estimate N from the measured mass (Fig. 1e) of the domain walls. In particular, if we consider a one-dimensional Ising model subject to a net transverse field Γ_c much less than the ferromagnetic coupling, the domain wall is a particle of mass

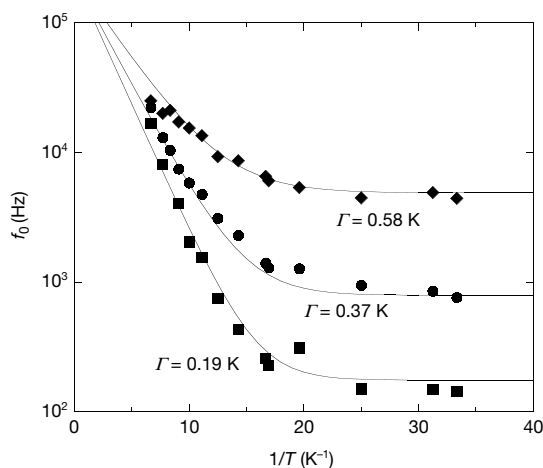


Figure 4 The characteristic frequency for magnetic domain relaxation as a function of both classical (T) and quantum (Γ) variables (data) along with the best fit of equation (5). At high temperature T , the relaxation is thermally activated over energy barriers that

decrease with increasing transverse field Γ . Below $T \approx 0.1$ K, the system smoothly enters a temperature-independent tunnelling regime of simple barrier tunnelling character. Error bars are comparable to the symbol sizes.

$m = \hbar^2/2a^2\Gamma_e$, where a is the lattice spacing. (This may be obtained from the semi-classical form $d^2E/dk^2 = \hbar^2/m$ applied to the domain wall dispersion of the transverse Ising model²⁰, and is the quantum spin-1/2 limit of the Döring mass^{21,22}.) For an array of N parallel chains the mass will be N times larger, that is $m = N\hbar^2/2a^2\Gamma_e$. Using the measured masses shown in Fig. 1e, we estimate that wall segments containing $N = 10$ spins tunnel together, and conclude that quantum relaxation in $\text{LiHo}_{0.44}\text{Y}_{0.56}\text{F}_4$ is coherent on the nanometre scale. □

Methods

We suspended a needle-shaped cylinder of $\text{LiHo}_{0.44}\text{Y}_{0.56}\text{F}_4$ (with aspect ratio 8 to minimize demagnetization effects) from the mixing chamber of a helium dilution refrigerator into the bore of an 8-T superconducting magnet oriented perpendicular to the crystalline c axis (to within 0.5°). A trim coil along the Ising direction nulled any unwanted longitudinal field component. The ordered state was always entered by cooling in large transverse field Γ and zero longitudinal field to the target temperature, and then reducing Γ through the phase boundary. Static measurements (that is, data of Fig. 2) were obtained using $200\text{ }\mu\text{m} \times 80\text{ }\mu\text{m}$ thin-film InAs Hall probes, crafted for low-temperature use²³, placed perpendicular to the Ising axis on the end of the sample cylinder. The dynamic response was measured after 12 h of equilibration time through the complex a.c. susceptibility, $\chi(f) = \chi'(f) + i\chi''(f)$, along the Ising axis with a standard gradiometer configuration, using digital lock-in amplifiers for the reference and signal channels. The energy splitting Γ between the originally degenerate Ising doublet (equation (1)) is calculated from the laboratory transverse magnetic field $H_\perp$ using the known crystal field levels of the Ho^{3+} ion²⁴.

Received 22 December 2000; accepted 16 August 2001.

- Leggett, A. J. *et al.* Dynamics of the dissipative two-state system. *Rev. Mod. Phys.* **59**, 1–85 (1987).
- Chudnovsky, E. M. & Gunther, L. Quantum tunneling of magnetization in small ferromagnetic particles. *Phys. Rev. Lett.* **60**, 661–664 (1988).
- Stamp, P. C. E., Chudnovsky, E. M. & Barbara, B. Quantum tunneling of magnetization in solids. *Int. J. Mod. Phys. B* **6**, 1355–1473 (1992).
- Braun, H.-B., Kyriakidis, J. & Loss, D. Macroscopic quantum tunneling of ferromagnetic domain walls. *Phys. Rev. B* **56**, 8129–8137 (1997).
- Hong, K. & Giordano, N. Evidence for domain wall tunneling in a quasi-one dimensional ferromagnet. *J. Phys. Condens. Matter* **8**, L301–L306 (1996).
- Copping, F. *et al.* Single domain switching investigated using telegraph noise spectroscopy: possible evidence for macroscopic quantum tunneling. *Phys. Rev. Lett.* **75**, 3513–3516 (1995).
- Wernsdorfer, W. *et al.* Macroscopic quantum tunneling of magnetization of single ferrimagnetic nanoparticles of barium ferrite. *Phys. Rev. Lett.* **79**, 4014–4017 (1997).
- Awschalom, D. D., Smyth, J. F., Grinstein, G., DiVincenzo, D. P. & Loss, D. Macroscopic quantum tunneling in magnetic proteins. *Phys. Rev. Lett.* **68**, 3092–3095 (1992).
- Friedman, J. R., Sarachik, M. P., Tejada, J. & Zio, R. Macroscopic measurement of resonant magnetization tunneling in high-spin molecules. *Phys. Rev. Lett.* **76**, 3830–3833 (1996).
- Thomas, L. *et al.* Macroscopic quantum tunneling of magnetization in a single crystal of nanomagnets. *Nature* **383**, 145–147 (1996).
- Barbara, B. *et al.* Quantum tunnelling in magnetic particles, layers and multilayers. *Phys. Scripta T* **49**, 268–273 (1993).
- Uehara, M. & Barbara, B. Noncoherent quantum effects in the magnetization reversal of a chemically disordered magnet: $\text{SmCo}_{0.5}\text{Cu}_{1.5}$. *J. Phys.* **47**, 235–238 (1986).
- Tejada, J., Zhang, X. X. & Chudnovsky, E. M. Quantum relaxation in random magnets. *Phys. Rev. B* **47**, 14977–14987 (1993).
- Vitale, S., Cavalleri, A., Cerdonio, M., Maraner, A. & Prodi, G. A. Thermal equilibrium noise with 1/f spectrum in a ferromagnetic alloy: Anomalous temperature dependence. *J. Appl. Phys.* **76**, 6332–6334 (1994).
- Arnaud, J. I., del Moral, A., de la Fuente, C., Ciria, M. & de Groot, P. A. J. Mesoscopic spin tunneling in the hard-random-axis-magnet amorphous alloy Tb_2Fe . *Phys. Rev. B* **50**, 547–550 (1994).
- Bitko, D., Rosenbaum, T. F. & Aeppli, G. Quantum critical behavior for a model magnet. *Phys. Rev. Lett.* **77**, 940–943 (1996).
- Barbara, B. Magnetization processes in high anisotropy systems. *J. Magn. Magn. Mater.* **129**, 79–86 (1994).
- Debye, P. *Polar Molecules* 91–94 (Chemical Catalogue, New York, 1929).
- Brooke, J., Bitko, D., Rosenbaum, T. F. & Aeppli, G. Quantum annealing of a disordered magnet. *Science* **284**, 779–781 (1999).
- Sachdev, S. *Quantum Phase Transitions* 39–46 (Cambridge Univ. Press, New York, 1999).
- Döring, W. Z. *Naturforsch.* **3A**, 373–379 (1948).
- Kyriakidis, J. & Loss, D. Bloch oscillations of magnetic solitons in anisotropic spin-1/2 chains. *Phys. Rev. B* **58**, 5568–5583 (1998).
- Pugel, E., Shung, E., Rosenbaum, T. F. & Watkins, S. P. Local magnetometry at high fields and low temperatures using InAs Hall sensors. *Appl. Phys. Lett.* **71**, 2205–2207 (1997).
- Hansen, P. E., Johansson, T. & Nevald, R. Magnetic properties of rare-earth fluorides: ferromagnetism in LiErF_4 and LiHoF_4 and crystal-field parameters at the rare-earth and Li sites. *Phys. Rev. B* **12**, 5315–5324 (1975).

Acknowledgements

We thank D. Bitko, S. Girvin, S. Nagel, P. Stamp and T. Witten for discussions. The work at the University of Chicago was supported primarily by the MRSEC Program of the National Science Foundation.

Correspondence and requests for materials should be addressed to T.F.R. (e-mail: t-rosenbaum@uchicago.edu).

Coexistence of superconductivity and ferromagnetism in URhGe

Dai Aoki*, Andrew Huxley*, Eric Ressouche*, Daniel Braithwaite*, Jacques Flouquet*, Jean-Pascal Brison†, Elsa Lhotel† & Carley Paulsen†

* CEA, Département de Recherche Fondamentale sur la Matière Condensée, SPSMS, 38054 Grenoble Cedex 9, France

† CNRS Centre de Recherche sur les Très Basses Températures, Grenoble 38042, France

The discovery¹ of superconductivity at high pressure (albeit over a restricted range) in the ferromagnetic material UGe_2 raised the possibility that bulk superconductivity might be found in other ferromagnets. The exact symmetry of the paired state and the dominant mechanism responsible for the pairing, however, remain unidentified. Meanwhile, the conjecture that superconductivity could occur more generally in ferromagnets has been fuelled by the recent observation of a low-temperature transition that suggests an onset of superconductivity in high-quality crystals of the itinerant-ferromagnet ZrZn_2 (ref. 2), although the thermodynamic signature of this transition could not be detected. Here we show that the ferromagnet URhGe is superconducting at ambient pressure. In this case, we find the thermodynamic signature of the transition—its form is consistent with a superconducting pairing of a spin-triplet type, although further testing with cleaner samples is needed to confirm this. The combination of superconductivity and ferromagnetism may thus be more common and consequently more important than hitherto realized.

Figure 1 shows the temperature dependence of the magnetization of a single crystal of URhGe . The material is ferromagnetic below a Curie temperature of $T_C = 9.5\text{ K}$, and has a low-temperature ordered moment of $\mu_s = 0.42\mu_B$ per formula unit, oriented along the c axis (URhGe is orthorhombic)^{3,4}. The quadratic low-temperature dependence of the squared magnetization is characteristic of simple itinerant ferromagnetism⁵, while neutron scattering reveals that the magnetization is almost entirely attributable to uranium 5f electrons. Contrary to the findings of a previous study¹, we find there to be no additional antiferromagnetic component to the order (see Methods). As for UGe_2 , the uniaxial anisotropy in the ferromagnetic state is known to be very large³ with an anisotropy field in excess of 100 T.

Although the space groups and detailed structures of URhGe and UGe_2 are different, both structures are orthorhombic and contain zigzag chains of nearest-neighbour uranium ions. In UGe_2 the nearest-neighbour uranium distance is $d_{U-U} = 3.8\text{ }\text{\AA}$ at zero pressure, but this is plausibly reduced at 1.3 GPa to $d_{U-U} = 3.5\text{ }\text{\AA}$ (due to a slight flattening of the chains⁶), a value almost identical to d_{U-U} in URhGe . In both materials the ordered moments lie in the plane of the zigzags, but they are aligned perpendicular to the chain axis in URhGe , and along this axis in UGe_2 . The physical properties of URhGe at zero pressure closely resemble those of UGe_2 at the high pressures (1.0–1.6 GPa) where superconductivity is found. In particular, the linear coefficient of the normal-state specific heat divided by temperature is large; $\gamma = 160\text{ mJ K}^{-2}\text{ mol}^{-1}$ for URhGe (ref. 7), compared with $\gamma = 120\text{ mJ K}^{-2}\text{ mol}^{-1}$ for UGe_2 (ref. 8). Therefore in both materials f electrons are implicitly involved in both highly correlated itinerant excitations and ferromagnetism. It was principally this quality and a favourable metallurgy that motivated the present study of URhGe . Even so, the behaviour of URhGe is not unique, as almost equivalent normal-state properties can be found in other ternary uranium compounds⁹.

Unlike the Curie temperature, which is insensitive to the quality of the samples examined, superconductivity was seen only in samples with a small normal-state residual electrical resistivity, ρ_0 . In particular, we found no sign of superconductivity down to 80 mK in samples with $\rho_0 \approx 30 \mu\Omega \text{ cm}$. For conventional superconductors there is no change in the phase of the order parameter with direction, and scattering from non-magnetic impurities hardly changes the superconducting transition temperature, T_s . But for non-conventional pairing the phase of the order parameter does change with direction, and superconductivity is expected to be completely destroyed when the electronic mean free path, l ,

inversely proportional to ρ_0 , is smaller than the coherence length, ξ . We therefore focus on measurements made on polycrystals which, to date, we have been able to synthesize with a lower residual resistivity ($\rho_0 \approx 2 \mu\Omega \text{ cm}$) than single crystals. In Fig. 2 the superconducting transitions for such samples are shown in the resistivity (ρ), alternating current susceptibility (χ_{ac}), specific heat (C), and static magnetization (M). The widths of the transitions at $T_s \approx 250 \text{ mK}$ are due in part to an inhomogeneous magnetic field distribution created by the division of the sample into ferromagnetic domains, although material variations could also account for some broadening. A simple estimate of l based on the measured quantities (γ , ρ_0 and H_{c2} , the upper critical field) confirm that these samples nevertheless comfortably satisfy the clean-limit condition, $\xi < l$. Several different samples of the same quality were studied, and all showed similar properties. The phase purity of the samples was verified by powder X-ray diffraction and electron microprobe analysis.

In zero applied field, χ_{ac} shows a clear diamagnetic response in the superconducting state, although the response is less than for an ideal Meissner phase, $\chi_{\text{Meissner}} = -1/(1-N)$ where N is the demagnetization factor determined from the sample geometry ($N \approx 0.1-0.2$). For URhGe even when the average magnetization of the sample is zero, the local field is not zero because each magnetic domain carries

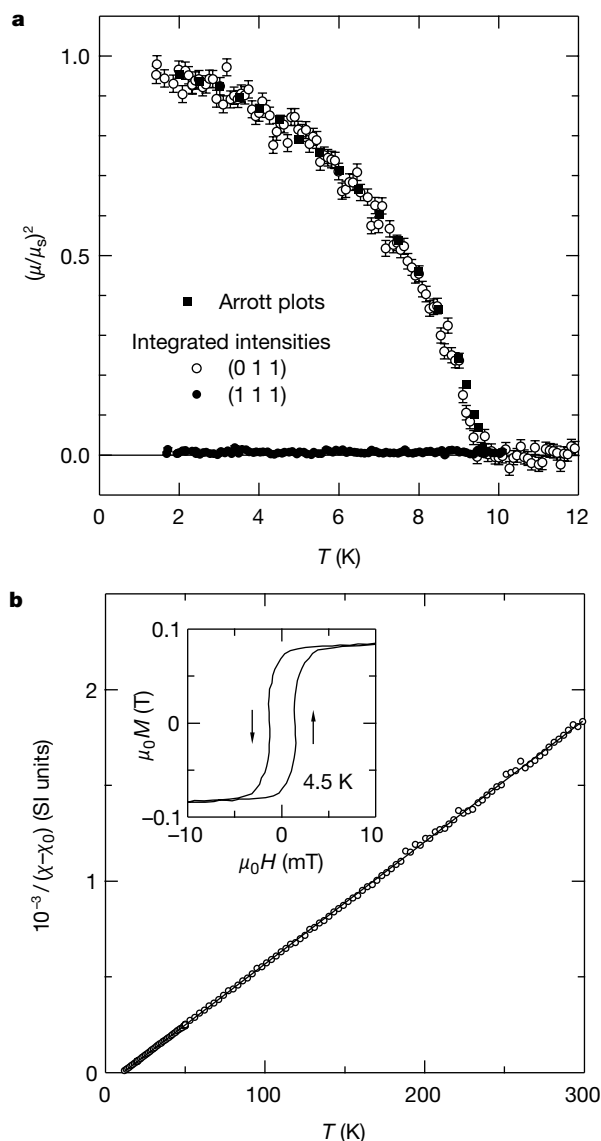


Figure 1 Magnetic properties of URhGe single crystals. **a**, The squared magnetization per formula unit versus temperature. The points determined from Arrott plots are normalized to a low-temperature limiting moment $\mu_s = 0.42 \mu_B$. The points proportional to the temperature-dependent part of the integrated intensity of a (011) neutron Bragg reflection are normalized to an almost identical value of μ_s . Also shown normalized to the same μ_s is the intensity of a (111) reflection, which is sensitive to an eventual antiferromagnetic component of the order. **b**, Temperature dependence of the easy axis susceptibility χ , plotted as $(\chi - \chi_0)^{-1}$, where $\chi_0 = 5 \times 10^{-4}$ (SI units) is a temperature-independent constant. The behaviour is similar to that predicted for itinerant ferromagnetism²⁰, with a Curie–Weiss moment of $1.8 \mu_B$ that is significantly larger than the ordered moment. Inset, the central part of a magnetic hysteresis loop at 4.5 K in the ferromagnetic state.

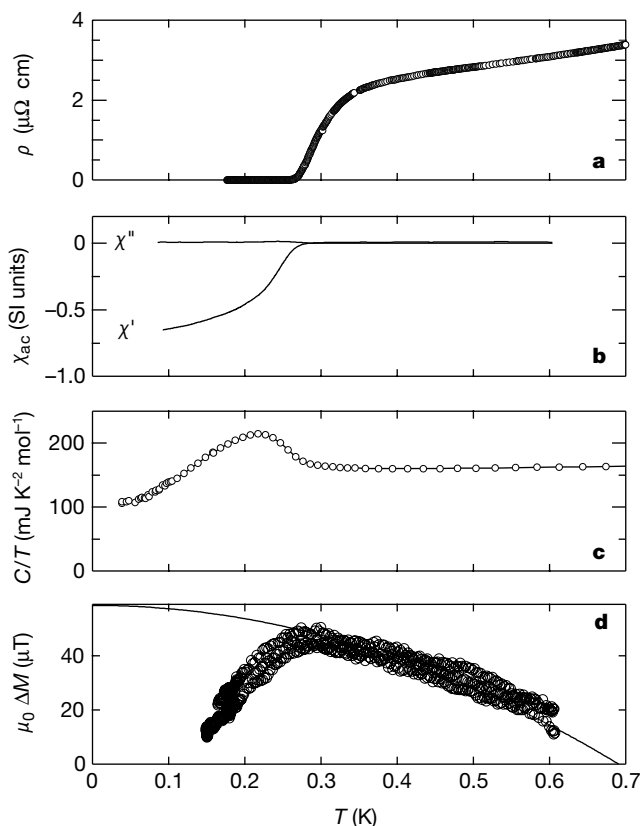


Figure 2 The superconducting transition measured in polycrystals of URhGe. **a–c**, The temperature dependence of the resistivity (**a**), the alternating current susceptibility (**b**), and the specific heat divided by the temperature (**c**), in zero applied field. **d**, The relative change of static magnetization in a constant applied field of 0.05 T. In **b**, both the in-phase (χ') and out of phase (χ'') components of the response are shown for a partially demagnetized sample. In **d**, the solid line shows the evolution of the relative magnetization calculated from $M = \bar{M}_0[1 - (T/T_0)^2]^{1/2}$, where $\bar{M}_0$ is chosen to give the measured absolute magnetization of the sample and T_0 is obtained from fitting this same dependence to the low-temperature data in Fig. 1a. The data points were taken after cooling the sample in a constant field, and are for both increasing and decreasing temperature.

a magnetization of the order of $\mu_0 M_s = 0.09$ T (corresponding to $\mu_s = 0.42\mu_B$ per formula unit). It is therefore probable that the sample is always in the mixed phase. A small change of field then penetrates the sample to a much greater depth than the London penetration depth¹⁰. This reduces the magnitude of χ_{ac} compared to the ideal Meissner phase value, and explains an observed weak dependence of χ_{ac} on applied field (not shown).

The observation of a large jump in the specific heat at the same temperature as the onset of diamagnetism, and close to the point where the resistivity becomes zero, clearly demonstrates that the transition to superconductivity is a bulk phase transition. It also shows that the superconductivity involves the same correlated itinerant electronic states that are responsible for the large γ .

Figure 2d shows the evolution of the static magnetization with temperature after the sample had been initially cooled in a constant field. For different fields, in the experimental range 0–0.1 T, similar changes of the magnetization were observed, but with magnitudes approximately proportional to the total magnetic moment of the sample. This confirms that the reversible behaviour seen at each field is due to a change in the intrinsic magnetization and not to changes of domain structure. The observed reversible temperature evolution of M above T_s follows $M = \bar{M}_0[1 - (T/T_0)^2]^{1/2}$ ($\bar{M}_0$ is determined from the measured total magnetization of the sample), with T_0 obtained from a fit to the low-temperature data in Fig. 1. But below T_s there is a significant decrease in the magnetization, not described by this dependence. Other measurements, in which the field was changed in the superconducting state (not shown), show that any flux-line pinning is very weak and can be neglected compared to the temperature evolution of M seen in Fig. 2d. In the absence of flux pinning, superconductivity is expected to reversibly expel flux giving rise to a change in the slope of M against T at T_s of magnitude $(\Delta M/dT) \approx (\bar{M}_0/M_s)(dH_{c2}/dT)/2\kappa^2$ ($\kappa \approx \lambda_L/\xi$ is the Ginzburg–Landau parameter and λ_L is the London penetration depth). This would imply that $\lambda_L \approx 9,000$ Å, which is larger

than the value observed¹¹ for the heavy-fermion superconductor UPt₃, even though the specific heat coefficient, γ , for URhGe is three times smaller. The change in slope that we detect, however, could be due to a combination of both flux expulsion and a small additional change in the magnetization. This additional change could come from either an intrinsic moment associated with the superconducting order, or a small modification of the underlying magnetic state: the opening of the superconducting gap would remove some low energy fluctuations, which might lead to a tiny increase in the ordered moment.

In Fig. 3 the temperature dependence of the upper critical field for superconductivity in a polycrystalline sample is shown. Due to the large width of the transition, H_{c2} is defined as the field at which the resistivity becomes zero. Even with this stringent definition, which underestimates the true H_{c2} , we get a rather large value of $\mu_0 H_{c2} = 0.71$ T at low temperature. As for UGe₂ near to the pressure where T_s is highest¹², this exceeds the weak-coupling paramagnetic limit applicable to conventional superconductors. As a much higher limiting field¹³ applies to spin-triplet superconductivity, this supports the identification of a spin-triplet state. Further, from the gradient dH_{c2}/dT near T_s , the Ginzburg–Landau coherence length can be estimated to be $\xi_{GL} \approx 180$ Å. The microscopic coherence length, ξ , is expected to be approximately equal to ξ_{GL} in an unconventional superconductor where l is necessarily larger than ξ . The small value of ξ indicates a large electronic effective mass—as expected, if the same electrons responsible for the large normal-state specific heat coefficient are involved in the superconductivity.

For three-dimensional ferromagnets, in the absence of spin–orbit coupling, and for a sufficiently large splitting between the majority and minority spin Fermi surfaces, the possibility of pairing opposite momentum states with anti-parallel spins is removed, as states with opposite spin are no longer degenerate. Triplet superconductivity is then the only channel available for pairing. In general a spin-triplet superconductor is characterized by three pairing amplitudes (or components of the order parameter), corresponding to the three symmetric combinations of spin, $\uparrow\uparrow + \downarrow\downarrow$, $\uparrow\downarrow$ and $\downarrow\uparrow$. A large spin splitting of the Fermi surfaces, as well as suppressing singlet pairing, may also suppress the first of these components. The splitting also creates an asymmetry between the remaining two amplitudes for $\uparrow\downarrow$ and $\downarrow\uparrow$, such that one of these amplitudes will be dominant, while the subdominant amplitude may disappear completely. Such a situation is realized in the A₁ phase of superfluid ³He, which exists over a very limited temperature range in a strong magnetic field. A similar state might also be anticipated in UGe₂ and URhGe with the quantization axis locked to the easy magnetization direction (although very strong spin–orbit interactions might lead to mixing between the different spin components).

Because URhGe belongs to the same symmetry class as UGe₂, the same classification scheme can be applied to the possible superconducting states in the presence of spin–orbit coupling¹⁴ or without spin–orbit coupling¹⁵. For both cases, if the pairing is indeed limited to one spin type, symmetry then requires the gap function to have a line node. For the specific heat, this would give a low-temperature limiting dependence of the form $C/T = \gamma_0 + aT$ and a smaller jump, ΔC , at T_s than for an isotropic superconductor¹⁶, in agreement with our experimental results. Even for non-conventional superconductors like UPt₃ and Sr₂RuO₄, C/T is predicted¹⁷ and observed^{18,19} to extrapolate to only a few per cent of γ at zero temperature. However, the present data for URhGe, like previous data for UGe₂ at high pressure⁸, give a finite zero-temperature intercept that in the present case is of the order of $\gamma/2$. Such a result, if confirmed to hold for cleaner samples, would provide strong evidence that the superconducting pairing indeed occurs for only one spin type. Electrons on Fermi surfaces of the opposite type would remain unpaired and continue to contribute to the specific heat, as in the normal state. This could also lead to a large value of λ_L . □

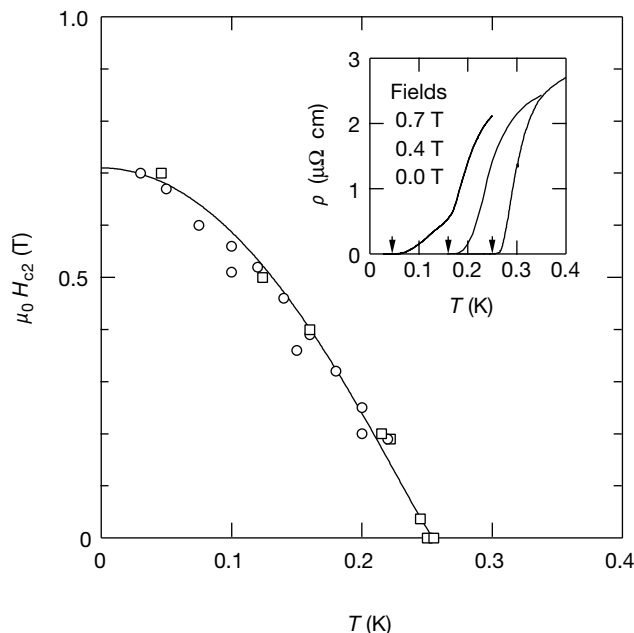


Figure 3 The upper critical field for superconductivity for a polycrystalline specimen of URhGe. The critical field at which the resistivity becomes zero ($\mu_0 H_{c2}$) is shown as a function of temperature. Circles, measurements made at constant temperature; squares, measurements made by changing the temperature in a constant field. The line is a guide to the eye. Inset, the temperature dependence of the resistivity at fields, from right to left, of 0, 0.4 and 0.7 T. Arrows indicate the temperatures at which the resistivity is zero.

Methods

The URhGe samples were manufactured from the constituent high-purity elements (depleted uranium 99.95%, rhodium 99.99% and germanium 99.9999%) with radio-frequency heating in water-cooled copper crucibles. The final synthesis was performed in a pressurized atmosphere of purified argon, followed by a subsequent anneal for 5 days under ultrahigh vacuum at 900 °C. The neutron study was carried out on the D23 CRG instrument at the Institute Laue Langevin. Around 300 reflections were collected from a single crystal both below and above T_C . The structural parameters evolve very slightly with temperature in this range. This evolution, however, gives rise to a much smaller change in the integrated Bragg intensities than the clear change in the integrated intensities due to ferromagnetic order (Fig. 1). For a powder sample Tran *et al.*⁴ previously reported an increase in the intensities of the (102) and (111) neutron reflections below T_C , of a magnitude comparable to the increase in the ferromagnetic (011) reflection. An increase in intensity of these two other peaks would not occur for a simple ferromagnetic structure, but would be present if the ordered state was instead a canted antiferromagnet as is allowed by symmetry (Tran *et al.* deduced an antiferromagnetic moment of $0.26\mu_B$ from their data). In our work we did not see any increase in intensity of the (102) and (111) reflections at low temperature (Fig. 1). We therefore conclude that any antiferromagnetic component of the order has a magnitude smaller than $0.06\mu_B$ compared to a ferromagnetic component of approximately $0.37\mu_B$.

The specific heat was measured by the relaxation method, and the magnetization and alternating current susceptibility were measured on a combination of purpose-built and commercial SQUID magnetometers. The measurements in a constant field, shown in Fig. 2d, were made without displacing the sample to avoid moving the sample through any slight field gradients. We note that the absolute values for the resistivity of the polycrystalline samples have been scaled to give approximately the directionally averaged resistivity at room temperature measured on single crystals. This procedure was necessary because the polycrystalline samples sometimes contained small cracks, which prevent the accurate assessment of their geometric factors needed to convert their resistances directly into units of resistivity.

Received 6 June; accepted 30 August 2001.

1. Saxena, S. S. *et al.* Superconductivity on the border of itinerant-electron ferromagnetism in UGe₂. *Nature* **406**, 587–592 (2000).
2. Pfeleiderer, C. *et al.* Coexistence of superconductivity and ferromagnetism in the *d*-band metal ZrZn₂. *Nature* **412**, 58–61 (2001).
3. de Boer, F. R., Brück, E., Sechovsky, V., Havela, L. & Buschow, K. H. J. UTX compounds in high magnetic fields. *Physica B* **163**, 175–178 (1990).
4. Tran, V. H., Troć, R. & André, G. Magnetic ordering in URhSi and URhGe. *J. Magn. Magn. Mater.* **186**, 81–86 (1998).
5. Lonzarich, G. G. & Taillefer, L. Effect of spin fluctuations on the magnetic equation of state of ferromagnetic or nearly ferromagnetic metals. *J. Phys. C* **18**, 4339–4371 (1985).
6. Huxley, A. *et al.* UGe₂: A ferromagnetic spin-triplet superconductor. *Phys. Rev. B* **63**, 144519–1–144519-13 (2001).
7. Hagmusa, I. H. *et al.* Magnetic specific heat of a URhGe single crystal. *Physica B* **281&282**, 223–225 (2000).
8. Tateiwa, N. *et al.* Pressure-induced superconductivity in a ferromagnet UGe₂. *J. Phys. Condens. Matter* **13**, L17–L23 (2001).
9. Sechovsky, V. & Havela, L. in *Handbook of Magnetic Materials* (ed. Buschow, K. H. J.) 1–290 (North-Holland, Amsterdam, 1998).
10. Campbell, A. M. The response of pinned flux vortices to low-frequency fields. *J. Phys. C* **2**, 1492–1501 (1969).
11. Broholm, C. *et al.* Anisotropic temperature dependence of the magnetic-field penetration in superconducting UPt₃. *Phys. Rev. Lett.* **65**, 2062–2065 (1990).
12. Sheikin, I. *et al.* Anisotropy and pressure dependence of the upper critical field of the ferromagnetic superconductor UGe₂. *Phys. Rev. B* (in the press).
13. Luk'yanchuk, I. A. & Mineev, V. P. Diamagnetic limit of superconductivity with triplet pairing. *JETP Lett.* **44**, 233–236 (1986).
14. Fomin, I. A. Symmetry of the order parameter in the UGe₂ superconductor. *JETP Lett.* 111–114 (2001).
15. Machida, K. & Ohmi, T. Phenomenological theory of ferromagnetic superconductivity. *Phys. Rev. Lett.* **86**, 850–853 (2001).
16. Leggett, A. J. A theoretical description of the new phases of liquid ³He. *Rev. Mod. Phys.* **47**, 331–414 (1975).
17. Hirschfeld, P., Vollhardt, D. & Wölfle, P. Resonant impurity scattering in heavy fermion superconductors. *Solid State Commun.* **59**, 111–115 (1986).
18. Brison, J. P. *et al.* Magnetism and superconductivity in heavy fermion systems. *J. Low Temp. Phys.* **95**, 145–152 (1994).
19. Nishizaki, S., Maeno, Y. & Mao, Z. Effect of impurities on the specific heat of the spin-triplet superconductor Sr₂RuO₄. *J. Low Temp. Phys.* **117**, 1581–1585 (1999).
20. Moriya, T. *Spin Fluctuations in Itinerant Electron Magnetism* (eds Cardona, M., Fulde, P. & Queisser, H. J.) Ch. 4 (Springer, Berlin, 1985).

Acknowledgements

We thank I. Fomin, V.P. Mineev and A. Wills for discussions, and P. Rodière, S. Sosin, R. Calemczuk, P. Lejay and F. Hardy for help in performing some of the reported measurements.

Correspondence and requests for materials should be addressed to A.H. (e-mail: huxley@cea.fr).

Atomic-scale imaging of insulating diamond through resonant electron injection

Kirill Bobrov, Andrew J. Mayne & Gérald Dujardin

Laboratoire de Photophysique Moléculaire, Bât. 210, Université Paris-Sud, 91405, Orsay, France

The electronic properties of insulators such as diamond are of interest not only for their passive dielectric capabilities for use in electronic devices¹, but also for their strong electron confinement² on atomic scales. However, the inherent lack of electrical conductivity in insulators usually prevents the investigation of their surfaces by atomic-scale characterization techniques such as scanning tunnelling microscopy (STM). And although atomic force microscopy could in principle be used, imaging diamond surfaces has not yet been possible. Here, we demonstrate that STM can be used in an unconventional resonant electron injection mode to image insulating diamond surfaces and to probe their electronic properties at the atomic scale. Our results reveal striking electronic features in high-purity diamond single crystals, such as the existence of one-dimensional fully delocalized electronic states and a very long diffusion length for conduction-band electrons. We expect that our method can be applied to investigate the electronic properties of other insulating materials and so help in the design of atomic-scale electronic devices.

In our experiment, a natural single crystal (100) of diamond was used. Such crystals are insulating because of their low dopant concentration. To prepare an atomically flat surface, the diamond sample was first hydrogenated *ex situ*³ in a hydrogen plasma. After introducing the sample into the ultrahigh-vacuum STM chamber (working pressure 2×10^{-11} torr), the adsorbed hydrogen was removed by *in situ* annealing⁴ at 1,100 °C. The hydrogenated surface can easily be imaged with the STM at negative sample bias⁵ because hydrogen is known to provide diamond with a p-type conductivity⁶. However, the hydrogen-free diamond surface is insulating; no tunnelling current could be obtained either at negative sample bias or at positive bias lower than about +4 V. This is indicated by the current–voltage (*I*–*V*) spectroscopy curves shown in Fig. 1, where the same behaviour is seen whatever the preset values of the tunnelling current. As a result, for any bias voltage between –6 V and +4 V, the STM tip crashed on the surface when trying to establish a tunnel current in the range 0.05–1 nA. If the clean diamond sample were not strongly insulating, we should be able to observe a tunnel current, because surface states are known to exist⁷ at about –3 V. A reliable quantitative measurement of such a low conductivity is very difficult as it should be done *in situ* under ultrahigh-vacuum conditions. Indeed, it is known that the diamond surface is easily contaminated by adsorption from the ambient air, causing dramatic changes in electrical conductivity⁸. Here, we found that the resistance of the $3.5 \times 2.5 \times 0.2$ mm diamond, sandwiched at each end between the molybdenum sample holder, increased from 10 kΩ when covered with hydrogen to 5 MΩ after removing the hydrogen by annealing. This is a crude measurement, which may be affected by some leakage current, and so represents only a lower limit of the sample resistance.

The absence of any tunnelling current in the –6 V to +4 V range is clearly related to the insulating character of the clean diamond surface and would *a priori* have prevented the use of the STM for any imaging. However, quite surprisingly, atomically resolved STM topographies of the clean diamond C(100)–(2 × 1) surface (the surface is parallel to the 100 crystallographic plane and in

orthogonal directions, the surface periodicity is respectively $2 \times$ and $1 \times$ the unit distance in the solid) could be recorded at very high sample bias ($U_{\text{bias}} = +5.9$ V) and constant current ($I = 1.1$ nA) as shown in Fig. 2a. In this image, terraces rotated by 90° are clearly visible. The periodic structure of bright and dark lines is observed on every terrace. The periodicity of ~ 0.5 nm, as measured from the scan profile shown in Fig. 2b, agrees well with the distance (0.504 nm) between the C–C dimer rows of the (2×1) reconstructed diamond surface⁸. Considering the atomic structure of the terraces close to the monoatomic steps (Fig. 2c), it can be seen that the bright areas in Fig. 2c are connected to each other, whereas the dark lines stand alone and never join each other. The interpretation of the STM topography (Fig. 2a) seems straightforward: the bright lines, joining at the step edges and visible as semicircular features, represent the inter-dimer row valleys whereas the dark lines correspond to the C–C dimer rows of the (2×1) reconstructed diamond surface.

One of the most striking features of the STM topography in Fig. 2a is that atomic resolution was obtained only at very high sample bias (+5.9 V), that is, above the diamond work function (5.3 V; ref. 9). This means that the STM operates in the near-field emission regime. Evidence of this is given by distance–voltage (z – V) spectroscopy (Fig. 3a–c), which shows the characteristic electron standing wave resonances^{10–12} located above the vacuum level of the diamond sample. The high sensitivity of these resonances to the electron–surface interaction has been recognized for a long time as a very powerful probe of the local electronic properties of surfaces^{10–12}. However, it had been generally accepted that the lateral spatial resolution was limited to rather large values^{10–12} of the order of 30 Å, although a better resolution seems to have been observed¹³ on Si(111). Therefore it is worthwhile to explain why atomic-scale resolution is obtained in the case of diamond.

A crucial observation is that atomically resolved STM topographies could be obtained only within a very narrow bias voltage range (± 0.1 V) matching almost exactly the energy of the lowest

standing-wave resonance. As the current is increased from 0.2 to 1.0 nA, the tip–surface distance is shortened by the STM feedback response and the electric field in the vacuum gap increases. The increased electric field modifies the shape of the potential barrier; the electronic potential well becomes narrower and the standing-wave resonances are shifted up in energy. As shown in Fig. 3d, the surface bias required for atomic-scale imaging follows the shift in the lowest standing wave resonance energy as the tunnel current is varied. Confirmation of this is obtained under nonresonance conditions, that is, when the sample bias did not match the first resonance energy, where only fuzzy structureless images were observed. This demonstrates that atomic resolution is obtained only through resonant electron tunnelling. We emphasize that this mode of STM imaging is markedly different from other methods previously used to image inhomogeneous and/or thin layers of insulators^{14,15}. These latter methods, based on the site-dependent electron transmission of thin insulator layers, could not be used here in the case of a thick insulating sample.

When tunnelling through the lowest standing-wave resonance, the surface bias has a value similar to the surface work function and

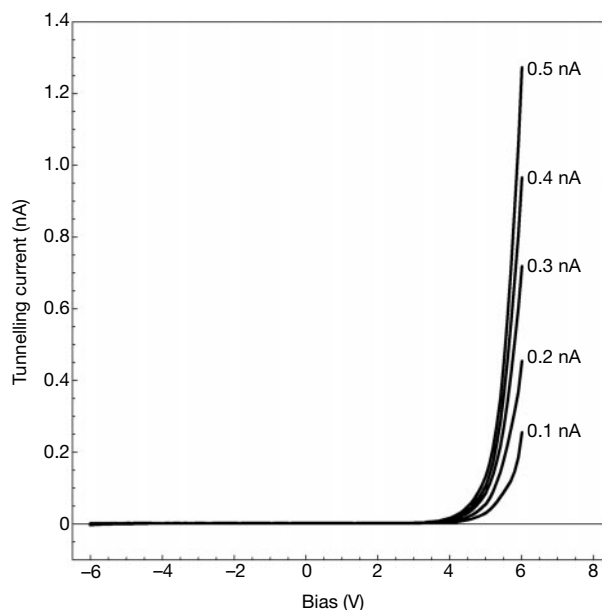


Figure 1 Current–voltage (I – V) spectroscopy of the clean C(100)– (2×1) surface as a function of tunnelling current. The I – V curves were recorded simultaneously with the scanning tunnelling microscope (STM) topography at $U_{\text{bias}} = +5.9$ eV. During normal imaging, the tunnel current (marked on the curve) is constant and gives an indication of the tip–sample separation. In I – V spectroscopy, the distance is fixed for each curve so that the current (y axis) varies as a function of V during the acquisition.

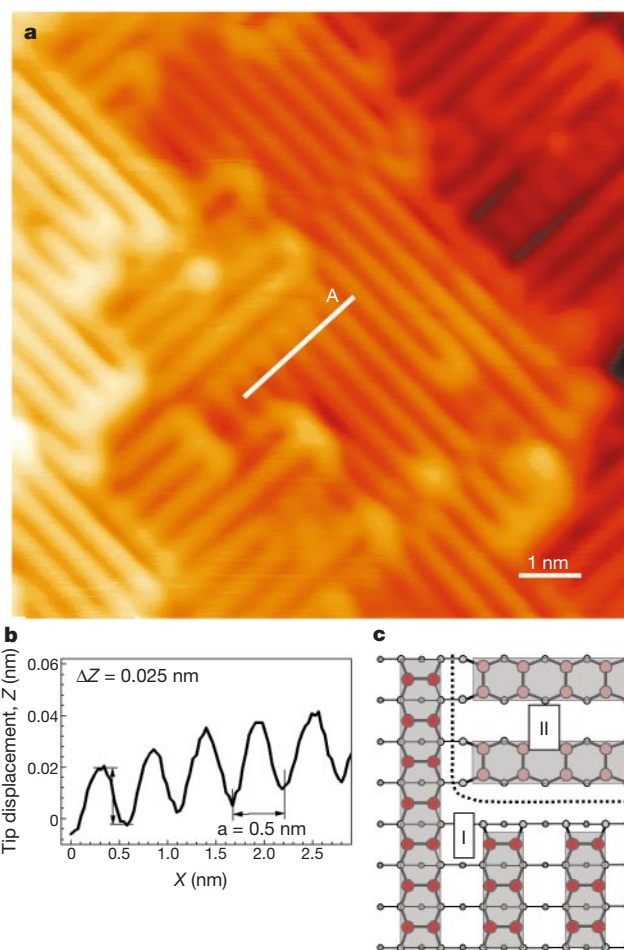


Figure 2 Clean diamond C(100)– (2×1) surface. **a**, The STM topography (10 nm $\times$ 10 nm) of the clean diamond surface recorded in the near-field emission regime ($U_{\text{bias}} = 5.9$ V, $I = 1.1$ nA). **b**, Height variation of the STM tip along the line A. **c**, Top-view of a monoatomic step on the two-domain (2×1) reconstructed surface. The coloured circles represent the carbon atoms belonging to the top four surface layers; the biggest circles represent the carbon–carbon dimers. The domains labelled as I and II represent the upper and lower terrace, respectively. The dimer rows are highlighted by shading, whereas the troughs between them are unfilled. The dashed line shows schematically the boundary between the domains.

electrons must tunnel through a long barrier (Fig. 4). This barrier acts as a filter transmitting only the electrons that have a wave vector $\mathbf{k}$ perpendicular to the surface. Similarly, the spreading of the tip–surface interaction is expected to be greatly reduced, that is, to atomic size as in the usual tunnel regime (when the surface bias is lower than the surface work function). However, this is not sufficient to explain the atomic resolution of the resonant electron tunnelling mode. To obtain atomic resolution, the energy of the standing-wave resonance needs to be corrugated at the atomic scale across the surface with a characteristic corrugation E large enough to be measurable. The measured z corrugation in the constant current mode is given by $\Delta z = dz/dV \Delta E$, where dz/dV is the slope of the $z(V)$ curve (see Fig. 3a) at the resonance.

An exceptional feature of the diamond surface, which explains the observed atomic resolution to a large extent, is the unusually large dz/dV slope at the first resonance (see Fig. 3a). At $20 \text{ \AA V}^{-1}$, it is an order of magnitude larger than in other materials^{10–12}. There are several fundamental reasons for this. The first is the very narrow width ($\sim 0.15 \text{ eV}$) of the first resonance. Such fineness is expected when the reflectivity of the electron wavefunction on the surface is high¹⁰. This is the case here, for two reasons. First, the bandgap of diamond is large, which places the bottom of the conduction band very close to the energy of the resonance (see Fig. 4). The large change in the $\mathbf{k}$ vector when the electrons penetrate into the diamond at the bottom of the conduction band results in this

high surface reflectivity. Second, the energy of the first resonance is only slightly above the energy of the conduction-band minimum (see Fig. 4). It follows that the off-resonance tunnel current should be very small compared to the on-resonance current, resulting in a large Δz step at the resonance. We emphasize that both effects, the narrow resonance width and the large Δz step, apply only to the resonance closest to the bottom of the conduction band. As a consequence, the fineness of the lowest resonance is much higher than that of the higher lying resonances.

The observed Δz corrugation of $0.25 \text{ \AA}$ (Fig. 2a and b) corresponds to a ΔE corrugation of about 12.5 mV . This means that the standing-wave resonance is 12.5 mV higher in energy when the tip is on top of a C–C dimer row than on top of an inter-dimer row. With reference to the properties of standing-wave resonances¹⁰ and image state resonances¹¹, we can begin to understand the origin of the ΔE corrugation, in terms of the atomic-scale nature of the electron scattering phenomenon on a surface. As schematically shown in Fig. 4, the standing-wave resonance energy depends on the shape of the potential energy well and the intensity and phase of electronic waves reflected^{12,16} from the diamond surface. The potential energy well is determined by the electrostatic potential, the image potential and at short distances the surface potential. It has been shown^{10,11} that any change in one of these potentials affects the energy of the resonances. We can therefore expect the ΔE corrugation to relate to atomic-scale variations of the image potential or the surface potential. We note that the apparent lack of resolution along the C–C dimer rows (see Fig. 2) has some important consequences. Indeed, referring to the height variation of the tip along the line A (Fig. 2b), the lateral STM resolution can be estimated to be better than $0.1 \text{ \AA}$. The lack of corrugation along the dimer rows indicates that the resonance energy, that is, the electron–surface interaction, although it is strongly anisotropic, shows no variation along the

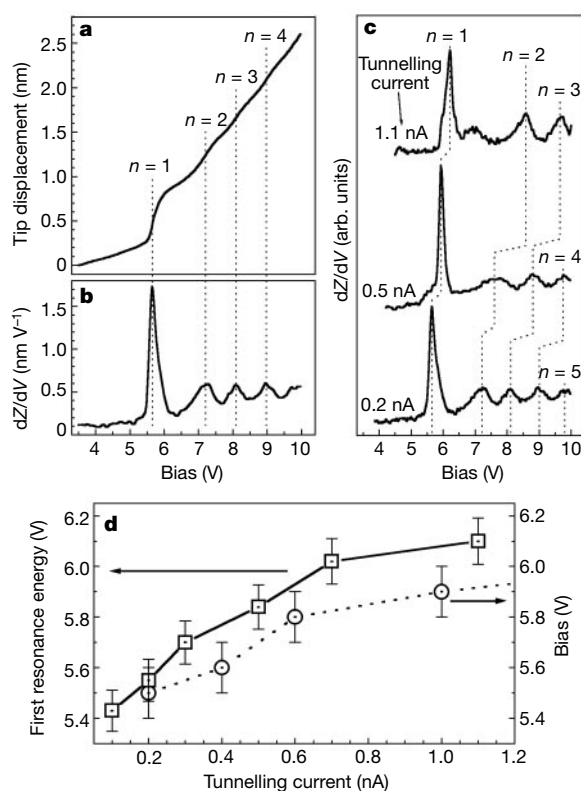


Figure 3 Distance–voltage z – V spectroscopy of the clean diamond C(100)– (2×1) surface. **a**, A typical spectroscopy curve recorded at $I = 0.2 \text{ nA}$ in the 3.5 – 10.0 V bias range; n is the resonance number. **b**, The numerical derivative dz/dV of the z – V spectroscopy curve shown in **a**. The energies of the barrier resonances (dotted lines) were taken at the maxima on the dz/dV curve. **c**, dz/dV curves as a function of the tunnelling current. The shift of the barrier resonance towards higher energies is shown by the dotted lines. **d**, The correlation between the bias at which atomically resolved STM topographies were obtained (dotted circles with error bars) and the first barrier resonance energy (dotted squares with error bars) as a function of the tunnelling current.

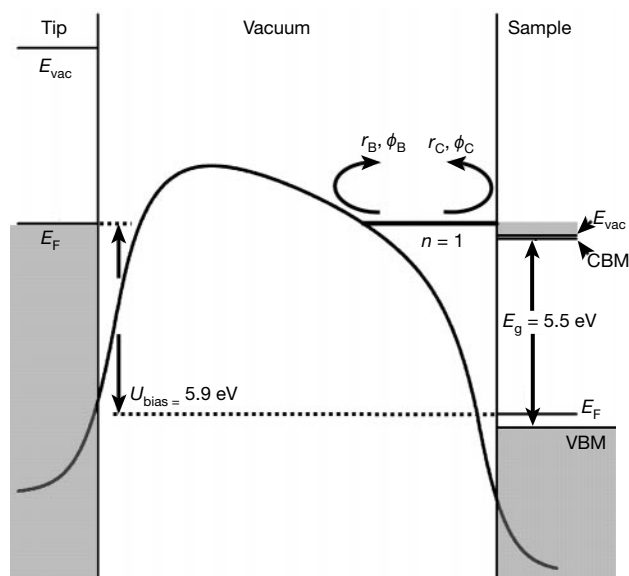


Figure 4 Potential well for the electrons scattered on the diamond surface. The potential barrier in the vacuum gap was constructed assuming a $10 \text{ \AA}$ tip–sample separation and a $0.4 \text{ V \AA}^{-1}$ electrostatic field in the vacuum gap. The actual shape of the potential barrier was obtained by a superposition of the electric field and the image potential of the surface, modified to include the surface potential contribution. For both the tip and the sample, the positions of the vacuum level, E_{vac} , and the Fermi level, E_F , are indicated. In addition, the surface gap E_g , the positions of the valence band maximum, VBM, and the conduction band minimum, CBM, are labelled. In the potential well, the reflection coefficients, r_B , r_C , and the dephasing, ϕ_B , ϕ_C , of the barrier and the diamond surface respectively, are noted at the first resonance, $n = 1$.

dimer rows. This is most probably due to the overlapping of the antibonding orbitals of C–C dimers⁸. It follows that the surface electronic image states¹⁷ can be considered to be delocalized one-dimensional (1D) electronic states¹⁸ along the dimer rows.

Finally, it is puzzling that we were able to transmit relatively high electron currents (1 nA) through a thick (0.2 mm) insulating diamond single crystal without any noticeable charging effect. This may be explained by an outstanding property of diamond single crystals; very long electron diffusion lengths of 150–250 μm have been measured¹⁹. Thus, diamond is insulating because of the lack of charge carriers, and yet electrons injected via the STM tip into the conduction band can be transported over very long distances without being trapped by defects.

We have used resonance electron tunnelling to image insulating diamond surfaces and to probe their electronic properties at the atomic scale using the STM. This approach could be used with other insulating materials provided they also have similar properties, that is, a large electronic band gap and long diffusion length of electrons in the conduction band. With these new capabilities for imaging insulating diamond surfaces at the atomic scale, surprising electron transport properties have been inferred, such as in 1D transport through the surface image states of dimer rows or in the injection of electrons in the conduction band. These results demonstrate that insulators like diamond could be advantageously used, beyond their well known passive dielectric properties, for active atomic-scale electronic devices. $\square$

Received 21 March; accepted 13 August 2001.

- Strahberger, C. & Vogl, P. Model of room-temperature resonant-tunneling current in metal/insulator and insulator/insulator heterostructures. *Phys. Rev. B* **62**, 7289–7296 (2000).
- Kohn, W. Theory of the insulating state. *Phys. Rev. A* **133**, 171–181 (1964).
- Thoms, B. D., Owens, M. S. & Butler, J. E. Production and characterization of smooth, hydrogen-terminated diamond C(100). *Appl. Phys. Lett.* **65**, 2957–2959 (1994).
- Bobrov, K., Schechter, Folman, M. & Hoffman, A. Deuterium adsorption-desorption from diamond (100) single crystal surfaces studied by TPD. *Diamond Related Mater.* **7**, 170–176 (1998).
- Nutzenadel, C. *et al.* STM investigations with atomic resolution on the (2 $\times$ 1) monohydride natural doped diamond (100) surface. *Surf. Sci.* **369**, L111–L116 (1996).
- Maier, F., Riedel, M., Mantel, B., Ristein, J. & Ley, L. Origin of surface conductivity in diamond. *Phys. Rev. Lett.* **85**, 3472–3475 (2000).
- Graupner, R. *et al.* Dispersion of surface states on diamond (100) and (111). *Phys. Rev. B* **55**, 10841–10847 (1997).
- Furtmüller, J., Hafner, J. & Kresse, G. Dimer reconstruction and electronic surface states on clean and hydrogenated diamond (100) surfaces. *Phys. Rev. B* **53**, 7334–7351 (1996).
- Diederich, L., Kuttel, O. M., Aebi, P. & Schlappach, L. Electron affinity and work function of differently oriented and doped diamond surfaces determined by photoelectron spectroscopy. *Surf. Sci.* **418**, 219–239 (1998).
- Becker, R. S., Golovchenko, J. A. & Swartzentruber, B. S. Electron interferometry at crystal surfaces. *Phys. Rev. Lett.* **55**, 987–990 (1985).
- Binnig, G. *et al.* Tunneling spectroscopy and inverse photoemission: image and field states. *Phys. Rev. Lett.* **55**, 991–994 (1985).
- Caamano, A. J. *et al.* Field emission interferometry with the scanning tunneling microscope. *Surf. Sci.* **426**, L420–L425 (1999).
- Kubby, J. A., Wang, Y. R. & Greene, W. J. Fabry-Perot transmission resonances in tunneling spectroscopy. *Phys. Rev. B* **43**, 9346–9349 (1991).
- Viernow, J. *et al.* Chemical imaging of insulators by STM. *Phys. Rev. B* **59**, 10356–10360 (1999).
- Libuda, J. *et al.* Structure and defects of an ordered alumina film on NiAl(110). *Surf. Sci.* **318**, 61–73 (1994).
- Radny, M. Electronic surface states within a new phase-shift multiple-scattering Green function approach. *J. Phys. C* **3**, 5525–5541 (1991).
- Höfer, U. *et al.* Time resolved coherent photoelectron spectroscopy of quantized electronic states on metal surfaces. *Science* **277**, 1480–1482 (1997).
- Silkin, V. M., Chulkov, E. V. & Echenique, P. M. Unusual dispersion of image potential states on the Be(1010) surface. *Phys. Rev. B* **60**, 7820–7823 (1999).
- Ristein, J., Stein, W. & Ley, L. Defect spectroscopy and determination of the electron diffusion length in single crystal diamond by total photoelectron yield spectroscopy. *Phys. Rev. Lett.* **78**, 1803–1806 (1997).

Acknowledgements

We thank P. Bergonzo and A. Hoffman for providing facilities to hydrogenate the diamond samples. This work is supported by the AFIRST programme, the European TMR network “Manipulation of individual atoms and molecules”, and the European IST-FET “Bottom-up-Nanomachines”, (BUN) programme.

Correspondence and requests for materials should be addressed to G.D. (e-mail: gerald.dujardin@ppm.u-psud.fr).

Selective assembly on a surface of supramolecular aggregates with controlled size and shape

Takashi Yokoyama*, Shiyoshi Yokoyama†, Toshiya Kamikado†, Yoshishige Okuno† & Shinro Mashiko†

* National Institute for Materials Science, 2268-1 Shimo-shidami, Morioka-ku, Nagoya 463-0003, Japan

† Communications Research Laboratory, 588-2 Iwaoka, Nishi-ku, Kobe 651-2401, Japan

The realization of molecule-based miniature devices with advanced functions requires the development of new and efficient approaches for combining molecular building blocks into desired functional structures, ideally with these structures supported on suitable substrates^{1–4}. Supramolecular aggregation occurs spontaneously and can lead to controlled structures if selective and directional non-covalent interactions are exploited. But such selective supramolecular assembly has yielded almost exclusively crystals or dissolved structures⁵; the self-assembly of adsorbed molecules into larger structures^{6–8}, in contrast, has not yet been directed by controlling selective intermolecular interactions. Here we report the formation of surface-supported supramolecular structures whose size and aggregation pattern are rationally controlled by tuning the non-covalent interactions between individual adsorbed molecules. Using low-temperature scanning tunnelling microscopy, we show that substituted porphyrin molecules adsorbed on a gold surface form monomers, trimers, tetramers or extended wire-like structures. We find that each structure corresponds in a predictable fashion to the geometric and chemical nature of the porphyrin substituents that mediate the interactions between individual adsorbed molecules. Our findings suggest that careful placement of functional groups that are able to participate in directed non-covalent interactions will allow the rational design and construction of a wide range of supramolecular architectures adsorbed to surfaces.

A large number of different selective and directional intermolecular connections has been developed in supramolecular chemistry⁵. Here we use a cyanophenyl substituent to control molecular aggregation, because it has a simple and symmetric structure as well as an asymmetric charge distribution at the cyano group that should introduce dipole–dipole interactions

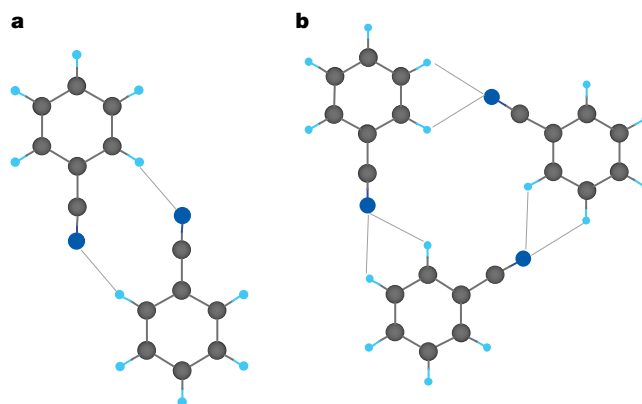


Figure 1 Calculated molecular aggregations. **a**, Cyanobenzene dimer; **b**, trimer. Colours correspond to the elements: H, light blue; C, black; N, dark blue.

between neighbouring cyanophenyl substituents. An illustration of this interaction is given in Fig. 1, which shows the optimized arrangement of a cyanobenzene dimer and trimer obtained in *ab initio* molecular orbital calculations at the MP2/6-31G* level. In the dimer, the cyano groups have an antiparallel configuration, whereas the trimer structure results from a cyclic arrangement of the cyano groups. The length of the CH...NC contacts is about 2.39 Å and 2.65 Å for the dimer and the trimer, respectively, and thus is shorter than the van der Waals distance of about 2.7 Å. The interaction energies are estimated to be $-7.12 \text{ kcal mol}^{-1}$ and $-12.40 \text{ kcal mol}^{-1}$ for the dimer and trimer structures, respectively, and thus comparable to the energy of hydrogen-bonding interactions. The relative orientation of molecules in these aggregates is therefore likely to be influenced by long-range dipole-dipole interaction, with hydrogen-bonding interactions further stabilizing the structure.

The characteristic intermolecular interaction between cyanophenyl groups is exploited to influence the aggregation of adsorbed porphyrin molecules. The porphyrin synthesized in this study is based on 5,10,15,20-tetrakis-(3,5-di-tertiarybutylphenyl) porphyrin ($\text{H}_2\text{-TBPP}$), which has a free-base porphyrin core and four di-tertiarybutylphenyl (tBP) substituents (Fig. 2a)⁹⁻¹³. To ensure selective molecular aggregation, we used a gold (111) as the substrate surface because of its inertness and its properties of reconstruction (see below). The surface was prepared by repeated cycles of Ar^+ sputtering and annealing in an ultrahigh-vacuum (UHV) chamber, followed by deposition of porphyrin molecules at room temperature by sublimation from a Knudsen cell in UHV. The resultant molecular arrangements were characterized by low-temperature scanning tunnelling microscopy (STM) carried out at 63 K (ref. 14).

Figure 3a shows an STM image of the Au(111) surface after a small amount of $\text{H}_2\text{-TBPP}$ was deposited, with isolated single molecules located at elbows that appear in the surface-reconstructed pattern of the Au substrate. The reconstruction of the Au(111) surface is known to include 'herringbone' patterns¹⁵, and a preferential nucleation of adsorbates at the elbows of this pattern has been

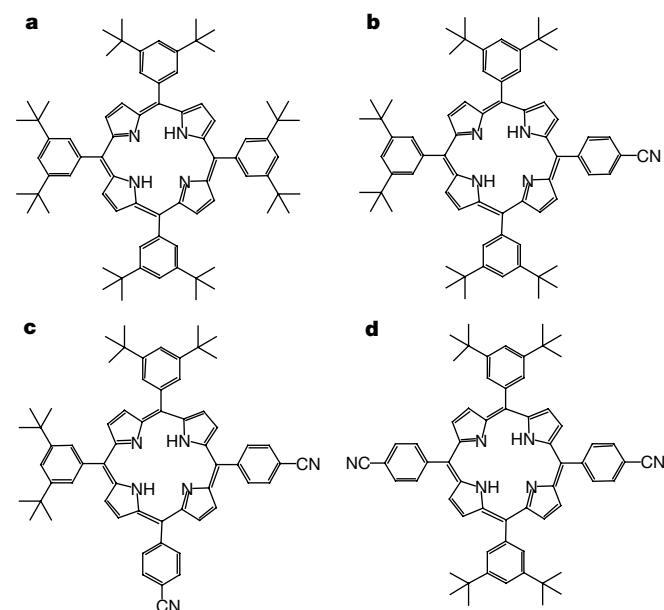


Figure 2 Structural formula of the porphyrins. **a**, $\text{H}_2\text{-TBPP}$, which is composed of a free-base porphyrin and four tBP substituents. In the ideal conformation, the phenyl rings are rotated by about 90° with respect to the porphyrin plane. **b**, CTBPP, where a tBP substituent of $\text{H}_2\text{-TBPP}$ is replaced with a cyanophenyl substituent. **c**, *cis*-BCTBPP, where two cyanophenyl groups were substituted at the *cis* position. **d**, *trans*-BCTBPP, where two cyanophenyl groups were substituted at the *trans* position.

reported for various systems^{8,16}. A high-resolution STM image (see Fig. 3e) shows that the individual $\text{H}_2\text{-TBPP}$ molecules are composed of four paired lobes surrounding two oblong protrusions. On the basis of the molecular dimensions, each lobe was assigned as one of the tertiarybutyl substituents, and the oblong protrusions as the central porphyrin ring. The appearance of the paired lobes suggests that the molecular plane of the phenyl rings is oriented to be closely aligned with the porphyrin mean plane, whereas about 90° rotations of the phenyl rings are expected for the ideal conformation of $\text{H}_2\text{-TBPP}$ (refs 9-11). We estimate the dihedral angle between the porphyrin and phenyl rings to be about 20° , which results from the rotational flexibility of the porphyrin-phenyl bonds and from the optimum adsorption of the tertiarybutyl groups onto the Au(111) surface. Upon further deposition, the $\text{H}_2\text{-TBPP}$ molecules exhibit a close-packed arrangement (Fig. 3e and i), which results from van der Waals interactions between the tBP substituents. (Further details of the adsorption of $\text{H}_2\text{-TBPP}$ molecules on the Au(111) surface are given in ref. 17.)

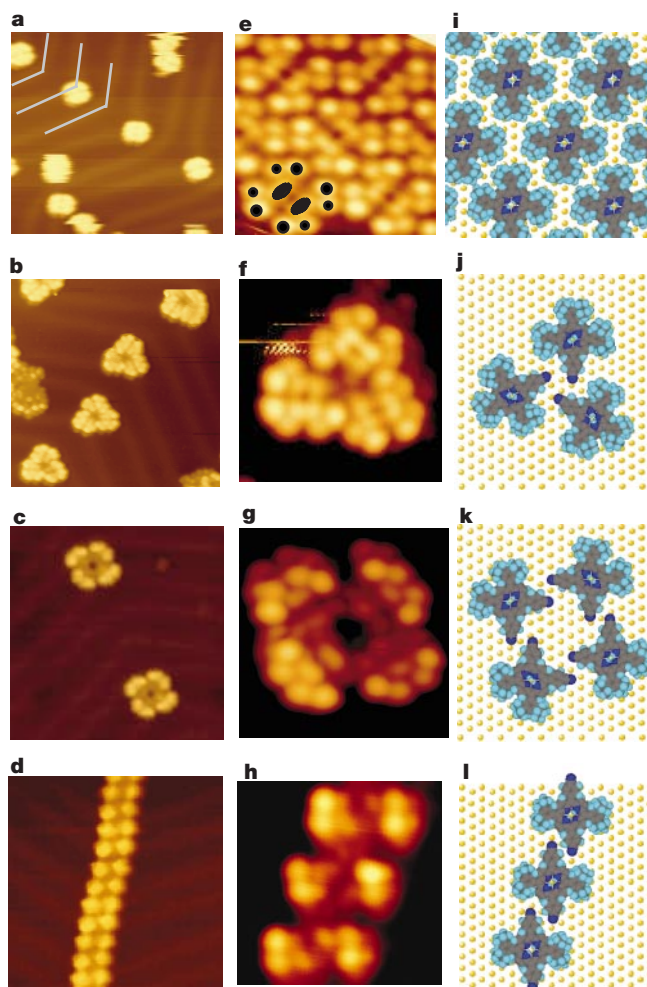


Figure 3 Scanning tunnelling microscope images at 63 K of supramolecular aggregations induced by the cyano groups. At small coverage, individual molecules or clusters are preferentially positioned at the elbows of the herringbone patterns of the Au(111) surface (imaged area, $20 \times 20 \text{ nm}$): **a**, $\text{H}_2\text{-TBPP}$; **b**, CTBPP; **c**, *cis*-BCTBPP; **d**, *trans*-BCTBPP. High-resolution STM image (imaged area, $5.3 \times 5.3 \text{ nm}$) and the molecular model: **e**, **i**, $\text{H}_2\text{-TBPP}$ island; **f**, **j**, CTBPP trimer; **g**, **k**, *cis*-BCTBPP tetramer; **h**, **l**, *trans*-BCTBPP wire. As indicated by black dots in **e**, the STM image of the single $\text{H}_2\text{-TBPP}$ molecule is composed of four paired lobes surrounding two oblong protrusions, corresponding to the tertiarybutyl substituents and the central porphyrin.

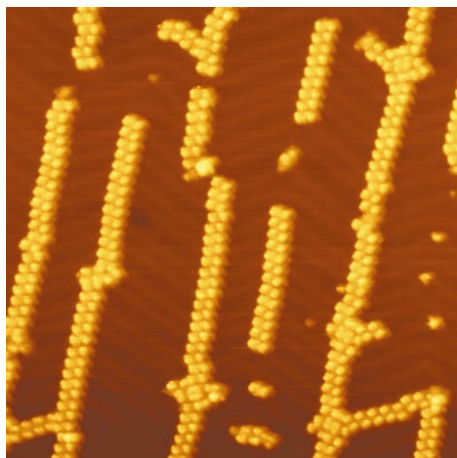


Figure 4 STM image at 63 K of *trans*-BCTBPP wires formed on the Au(111) surface (imaged area, 70 × 70 nm). The supramolecular wires extended across the elbows of the herringbone patterns, and the observed maximum length is above 100 nm, which depends on the terrace width.

To introduce the characteristic interaction of the cyano substituents, we have synthesized (cyanophenyl)-tris(di-tertiarybutylphenyl) porphyrin (CTBPP), where a tBP group of H₂-TBPP was replaced with a cyanophenyl substituent (Fig. 2b). At low coverage, most CTBPP molecules assemble into triangular clusters on the Au(111) surface. As shown in Fig. 3b, identical clusters are located separately at the elbows of the herringbone patterns. From the molecular structure of CTBPP, three paired lobes are expected as a single molecule in the STM image, because a paired lobe corresponds to one di-tertiarybutyl substituent. The high-resolution STM image of Fig. 3f shows that the triangular cluster is a CTBPP trimer. The cyanophenyl substituents are assembled into a cyclic configuration in the trimer structure of Fig. 3j, in agreement with the cyanobenzene trimer aggregation of Fig. 1b. Compared with the characteristic aggregation of CTBPP, we have not observed such supramolecular clusters for (phenyl)-tris(di-tertiarybutylphenyl) porphyrins, where the cyano substituent was synthetically removed from the CTBPP molecule. In this case, a random arrangement of the molecules was formed on the surface, confirming that the supramolecular aggregation is dominated by the cyano groups.

To probe the supramolecular aggregation in our system further, we substituted one more cyano group to give bis(cyanophenyl)-bis(di-tertiarybutylphenyl) porphyrin (BCTBPP), which forms two types of isomers (*cis* and *trans*) with respect to the configuration of two cyanophenyl substituents (Fig. 2c and d). Figure 3c shows that the *cis*-BCTBPP molecules are aggregated into a supramolecular tetramer. In this structure, the antiparallel intermolecular connections of all the cyanophenyl substituents lead to a macrocyclic arrangement of the porphyrin molecules, forming a molecular ring (see Fig. 3g and k).

In contrast to the macrocyclic clusters of CTBPP and *cis*-BCTBPP, sequential aggregation was achieved for the *trans*-BCTBPP molecules, where two cyanophenyl groups were substituted at the trans positions. Figure 3d and h shows the STM images of the Au(111) surface after submonolayer deposition of the *trans*-BCTBPP molecules. In this structure, the antiparallel configuration between the cyanophenyl substituents results in a linear arrangement of the *trans*-BCTBPP molecules, forming the supramolecular wires. As shown in the large-scan STM image of Fig. 4, most of the individual wires extended across the elbows of the herringbone patterns, because the molecules were initially nucleated at the elbows. Furthermore, it is remarkable that the maximum length of the

straight wires is above 100 nm, although some branches are also formed due to the three-fold symmetry of Au(111).

These selective aggregations of the porphyrin molecules result from the antiparallel or cyclic configurations of the cyano substituents relative to each other. From the molecular models and the high-resolution STM images, we estimated the length of the CH...NC contact between cyanophenyl substituents to be about 2.6 Å and 2.5 Å for the antiparallel and cyclic configurations, respectively. These values are almost identical to those for the calculated cyanobenzene aggregates, suggesting hydrogen-bonding interactions. Although the strength of the non-covalent intermolecular interactions in our supramolecular aggregates is thus significantly weaker than encountered in covalent bonds, electronic coupling across hydrogen-bonded interfaces can nevertheless allow electron or energy transfer¹⁸, as has been observed in biological systems^{19,20}. In principle, systems similar to those described here should therefore be capable of exhibiting useful electronic and optoelectrical functions. Furthermore, the central porphyrins and cyanophenyl substituents within the clusters or wires should be decoupled from both the substrate surface and surrounding clusters in our present systems, because the bulky tBP insulators attached to the surface prevent any direct interaction. This indicates that unusual chemical and physical properties associated with the size and shape of the aggregate cluster should remain, even after deposition on the surface. □

Received 30 April; accepted 1 August 2001.

- Wada, Y. *et al.* Prospects and problems of single molecule information devices. *Jpn J. Appl. Phys.* **39**, 3825–3846 (2000).
- Collier, C. P. *et al.* Electronically configurable molecular-based logic gates. *Science* **285**, 391–394 (1999).
- Lopinski, G. P., Wayner, D. D. M. & Wolkow, R. A. Self-directed growth of molecular nanostructures on silicon. *Nature* **406**, 48–51 (2000).
- Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. *Nature* **408**, 541–548 (2000).
- Lehn, J.-M. *Supramolecular Chemistry: Concept and Perspectives* (VCH, Weinheim, 1995).
- Furukawa, M., Tanaka, H., Sugiura, K., Sakaya, Y. & Kawai, T. Fabrication of molecular alignment at the specific sites on Cu(111) surface using self-assembly phenomena. *Surf. Sci. Lett.* **445**, L58–L63 (2000).
- Barth, J. V. *et al.* Building supramolecular nanostructures at surfaces by hydrogen bonding. *Angew. Chem. Int. Edn Engl.* **39**, 1230–1234 (2000).
- Bohringer, M. *et al.* Two-dimensional self-assembly of supramolecular clusters and chains. *Phys. Rev. Lett.* **83**, 324–327 (1999).
- Sugiura, K. *et al.* X-ray photoelectron spectroscopy of metallo porphyrins having bulky substituents: Standard values of core ionization potentials. *Chem. Lett.* 841–842 (1999).
- Jung, T. A., Schlittler, R. R., Gimzewski, J. K., Tang, H. & Joachim, C. Controlled room-temperature positioning of individual molecules: Molecular flexure and motion. *Science* **271**, 181–184 (1996).
- Jung, T. A., Schlittler, R. R. & Gimzewski, J. K. Conformational identification of individual adsorbed molecules with the STM. *Nature* **386**, 696–698 (1997).
- Moresco, F. *et al.* Conformational changes of single molecules induced by scanning tunneling microscopy manipulation: A route to molecular switching. *Phys. Rev. Lett.* **86**, 672–675 (2001).
- Moresco, F. *et al.* Low temperature manipulation of big molecules in constant height mode. *Appl. Phys. Lett.* **78**, 306–308 (2001).
- Yokoyama, T. & Takayanagi, K. Size quantization of surface-state electrons on the Si(001) surface. *Phys. Rev. B* **59**, 12232–12236 (1999).
- Barth, J. V., Brune, H., Ertl, G. & Behm, R. J. Scanning tunneling microscopy observations on the reconstructed Au(111) surface: Atomic structure, log-range superstructure, rotational domains, and surface defects. *Phys. Rev. B* **42**, 9307–9318 (1990).
- Chambliss, D. D., Wilson, R. J. & Chiang, S. Nucleation of ordered Ni island arrays on Au(111) by surface-lattice dislocations. *Phys. Rev. Lett.* **66**, 1721–1724 (1991).
- Yokoyama, T., Yokoyama, S., Kamikado, T. & Mashiko, S. Nonplanar adsorption and orientational ordering of porphyrin molecules on Au(111). *J. Chem. Phys.* **115**, 3814–3818 (2001).
- de Rege, P. J. F., Williams, S. A. & Therien, M. J. Direct evaluation of electronic coupling mediated by hydrogen bonds: implications for biological electron transfer. *Science* **269**, 1409–1413 (1995).
- Find, H.-W. & Schonenberger, C. Electrical conduction through DNA molecules. *Nature* **398**, 407–410 (1999).
- Cai, L., Tabata, H. & Kawai, T. Self-assembled DNA network and their electrical conductivity. *Appl. Phys. Lett.* **77**, 3105–3106 (2000).

Acknowledgements

T.Y. would like to thank K. Kimura, Y. Takaguchi and Y. Wakayana for useful discussions.

Correspondence and requests for materials should be addressed to T.Y. (e-mail: Yokoyama.takashi@nims.go.jp).

Acclimatization of soil respiration to warming in a tall grass prairie

Yiqi Luo, Shiqiang Wan, Dafeng Hui & Linda L. Wallace

Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma 73019, USA

The latest report by the Intergovernmental Panel on Climate Change (IPCC) predicts a 1.4–5.8 °C average increase in the global surface temperature over the period 1990 to 2100 (ref. 1). These estimates of future warming are greater than earlier projections, which is partly due to incorporation of a positive feedback. This feedback results from further release of greenhouse gases from terrestrial ecosystems in response to climatic warming^{2–4}. The feedback mechanism is usually based on the assumption that observed sensitivity of soil respiration to temperature under current climate conditions would hold in a warmer climate⁵. However, this assumption has not been carefully examined. We have therefore conducted an experiment in a tall grass prairie ecosystem in the US Great Plains to study the response of soil respiration (the sum of root and heterotrophic respiration) to artificial warming of about 2 °C. Our observations indicate that the temperature sensitivity of soil respiration decreases—or acclimatizes—under warming and that the acclimatization is greater at high temperatures. This acclimatization of soil respiration to warming may therefore weaken the positive feedback between the terrestrial carbon cycle and climate.

The third IPCC assessment on climate change has made a great advance in predicting future global climate with, among other improvements, coupled climate and carbon cycle models. Such coupled models are essential to examine biotic feedbacks into the climatic system. Climatic warming, on one hand, potentially stimulates nutrient mineralization and lengthens growing seasons, which consequently increases plant growth and carbon sequestration⁶ (Fig. 1). On the other hand, warming can accelerate biospheric metabolism, resulting in the greater release of heat-trapping gases to the atmosphere^{7–9}, which in return reinforces anthropogenic warming. Indeed, a coupled climate–carbon cycle model by Cox *et al.* (ref. 2) predicts 8.0 °C of global terrestrial warming by the year 2100 rather than the 5.5 °C predicted without the climate–carbon cycle connection. A major uncertainty associated with such a prediction is the assumption that the observed sensitivity of soil respiration to temperature under the current climate—approximately, a doubling of respiration for every 10 °C increase, that is, the temperature quotient $Q_{10} = 2$ —would hold in a future warmed climate⁵. Here we provide experimental evidence that this assumption may be unwarranted.

We have conducted a field warming experiment in a tall grass prairie in central Oklahoma, USA (34° 59' N, 97° 31' W) since 21 November 1999 to study respiratory sensitivity to climate change. We also used clipping to mimic hay harvesting, which is extensively practised in the US Great Plains, in an attempt to examine the interaction of global warming and land-use change. Experimental warming, on average, increased the daily mean air temperature by 1.1 °C, and daily mean soil temperature by 2.0 °C and 2.6 °C in the unclipped and clipped subplots, respectively (Fig. 2a and b).

With a fixed Q_{10} of 2, an increase of 2.0–2.6 °C in soil temperature should cause soil respiration to increase by 15–20%. However, our data showed that experimental warming generally caused no significant changes in soil respiration (Fig. 2e and f). Soil respiration in the warmed plots decreased by, on average, 5% in comparison to that in unwarmed plots without clipping and increased by 0.2% with clipping. This result is supported by long-term field studies in lowland tundra¹⁰ and Finnish soils¹¹, showing that measured soil

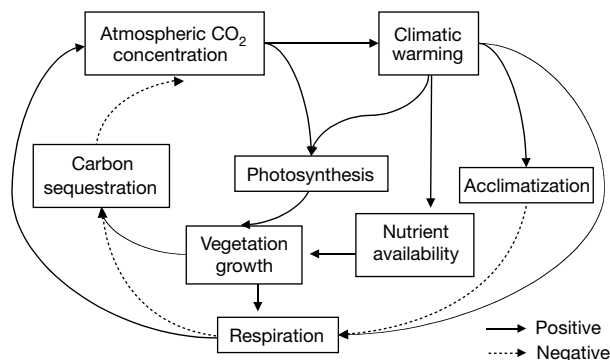


Figure 1 Schematic diagram of major feedbacks in a coupled climate–carbon cycle system. Climatic warming could trigger both a negative feedback through stimulation of nutrient mineralization, plant growth and carbon sequestration on one hand and a positive feedback via an increase in ecosystem respiration to release additional CO₂ on the other hand. Our study provides experimental evidence that acclimatization of soil respiration to warming could potentially weaken the positive feedback through the warming–respiration–atmospheric CO₂ connection. Those mechanisms of climate–carbon cycle coupling may operate at different timescales²⁸.

respiration is not significantly different between two or more temperature regimes. Initial increases in soil respiration under warming in other experiments^{12–14} declined over time^{15,16}. It was also observed that experimental warming reduced soil respiration in an alpine meadow study¹⁷.

To examine the temperature sensitivity of soil respiration, we conducted regression analyses (Fig. 3) using $R_s = ae^{bT}$, where R_s is soil respiration, T is soil temperature, coefficient a is the intercept of soil respiration when temperature is zero, and coefficient b represents the temperature sensitivity of soil respiration. In order to reduce variation in this analysis, we used the soil temperature that was measured at a depth of 5 cm at the time when soil respiration was measured. Four data points of soil respiration per treatment during the summer drought period (Fig. 2c and d) were excluded when soil moisture was below 7% (grams of water per 100 grams of soil, above which soil respiration was not significantly regulated by soil moisture (X. Liu, S.W., B. Su, D.H. & Y.L., unpublished work). Analysis of data from the four treatments (unwarmed, unclipped; warmed, unclipped; unwarmed, clipped; and warmed, clipped) indicates that warming did not significantly affect coefficient a with either clipping regime but did alter the temperature sensitivity coefficient b (Table 1). We merged two a values between unwarmed and warmed plots and then re-estimated the temperature sensitivity coefficient b . The difference in the re-estimated b values between warmed and unwarmed plots was found to be statistically significant with either clipping treatment (Table 1). The b values were used to calculate a respiration quotient $Q_{10} (= e^{10b})$, which decreased from 2.70 in the unwarmed, unclipped plots to 2.43 in the warmed, unclipped plots. The Q_{10} value decreased from 2.25 in the unwarmed, clipped plots to 2.10 in the warmed, clipped plots.

The decrease in temperature sensitivity of soil respiration under warming could result from several mechanisms, including concurrent reduction in plant production leading to less root respiration¹⁷, soil drying reducing root and microbial activity¹⁴, and substrate limitation^{13,16,18}. We measured aboveground live biomass on 28 April, 18 May and 18 November 2000, which was significantly higher (18.3% on average) in the warmed plots than in the unwarmed plots. Increased plant biomass growth presumably enhances root respiration, thus contributing to soil respiration, rather than reducing it.

Measured soil moisture was, on average, 6.4% lower in the warmed plots than in the unwarmed plots over the entire experimental period. To discern the effects of moisture from temperature

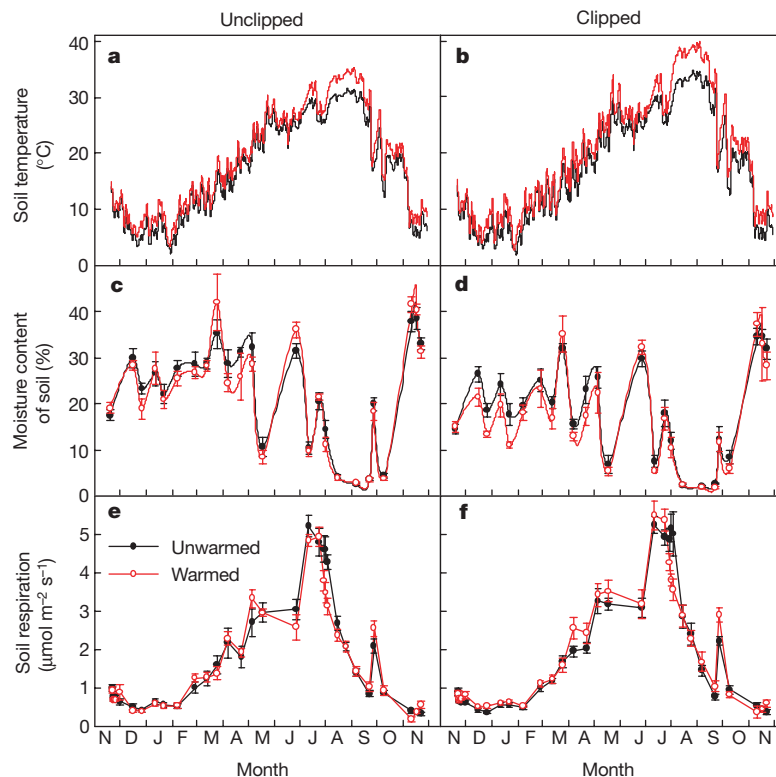


Figure 2 Time courses of measured soil temperature, soil moisture content, and soil respiration from November 1999 to November 2000. **a, b**, Soil temperature; **c, d**, soil moisture content; **e, f**, soil respiration. Data from unwarmed (black lines) and warmed (red lines) plots are shown without clipping (left column) and with clipping (right column). Soil moisture content and respiration are displayed as mean $\pm$ standard error with the sample

size $n = 5$. The sharp decreases in soil respiration in August and September were caused by severe summer drought. The slight increase under warming from April to July with clipping (**f**) was correlated with increased biomass growth. Overall, warming treatment did not significantly change soil respiration either with or without clipping.

on soil respiration, we used the Chi-square (χ^2) test of independence to examine whether effects of warming on soil respiration were correlated with warming-induced changes in soil moisture. We developed three contingency tables for unclipped plots, clipped plots and merged data from unclipped and clipped plots, respectively. The χ^2 tests indicate that the effects of warming on soil respiration are independent of warming-induced changes in soil moisture ($P = 0.31$ – 0.84 ; see the Supplementary Information). Thus, decreased temperature sensitivity of soil respiration to warming was unlikely to be caused by soil drying.

Substrate quality and quantity regulate responses of soil respiration to temperature^{13,18,19}, potentially resulting in acclimatization²⁰. That is, the temperature sensitivity of the soil ensemble decreases when exposed to a warmed environment. Respiratory acclimatization possibly results from changes in the composition of the microbial community²¹, reduced respiratory capacity in soil and/

or other physiological and ecological adjustments²² in response to a limited substrate supply. In our experimental site, soil carbon content is 1.42% on a mass basis, which is relatively low in comparison to other warming experimental sites¹⁵. Soil with low carbon content may be prone to acclimatization.

To evaluate the degree of acclimatization of soil respiration to experimental warming, we define a theoretical isocline of full acclimatization by $R_w = R_u$ at T_w , where R_w is the soil respiration in the warmed plots, R_u in the unwarmed plots, and T_w is the soil temperature in the warmed plots. Comparison of the temperature sensitivity curve in the warmed plots with the isocline indicates that soil respiration was less than fully acclimatized to warming when the temperature was lower than 13.9 °C and more than fully acclimatized when the temperature is higher than 13.9 °C in the unclipped subplots (Fig. 3a). When plants were clipped, soil respiration was less than fully acclimatized at temperatures lower than 20.8 °C and

Table 1 Statistical analysis of temperature relationships of soil respiration

Treatment		<i>a</i>	<i>b</i>	<i>r</i> ²	<i>t_a</i>	<i>t_b</i>	<i>Q</i> ₁₀
Unclipped	Unwarmed	0.267 $\pm$ 0.033	0.104 $\pm$ 0.005	0.967	–1.195	2.859*	–
Unclipped	Warmed	0.332 $\pm$ 0.070	0.085 $\pm$ 0.008	0.870			–
Clipped	Unwarmed	0.333 $\pm$ 0.058	0.086 $\pm$ 0.006	0.923	–1.089	2.074	–
Clipped	Warmed	0.430 $\pm$ 0.112	0.070 $\pm$ 0.009	0.767			–
Pooled coefficient <i>a</i> between two warming treatments							
Unclipped	Unwarmed	0.300 $\pm$ 0.033	0.099 $\pm$ 0.001	0.966	–	7.824**	2.70
Unclipped	Warmed		0.089 $\pm$ 0.002	0.868			2.43
Clipped	Unwarmed	0.381 $\pm$ 0.065	0.081 $\pm$ 0.001	0.921	–	4.164**	2.25
Clipped	Warmed		0.074 $\pm$ 0.002	0.766			2.10

* indicates a statistical significance at $P < 0.01$; and ** at $P < 0.001$.

a and *b* are two coefficients in the regression line $R_s = ae^{bT}$, where R_s is soil respiration and T is soil temperature. r^2 is the determinant coefficient, t_a and t_b are the Student *t* values for testing statistical significance in coefficient *a* and *b* values, respectively, between unwarmed and warmed plots in each clipping regime. Coefficient *a* is not statistically different between the two warming levels, so we pooled *a* values together and then re-estimated *b* for each of the treatments. *Q*₁₀ is the temperature quotient ($= e^{10b}$).

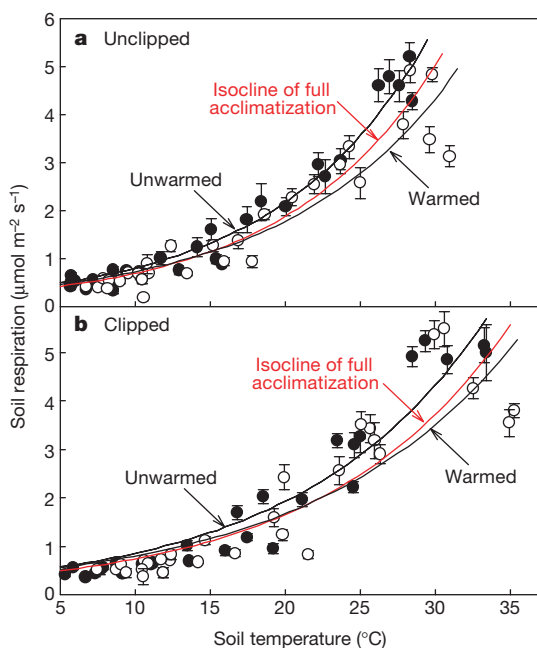


Figure 3 The relationships between soil respiration and temperature in unwarmed (filled circles) and warmed (open circles) treatments with standard errors. **a**, Unclipped subplots; **b**, clipped subplots. Response curves of soil respiration to temperature in the warmed treatment are below that in the unwarmed treatment, indicating that warming reduced temperature sensitivity. The theoretical isoclines of full acclimatization are defined as identical rates of respiration in the warmed plots to those in the unwarmed plots, at the temperature in the warmed plots. Soil temperature used in this analysis was measured at a depth of 5 cm at the time when soil respiration was measured. It was 1.5 °C and 1.9 °C higher in the warmed than the unwarmed plots without and with clipping, respectively.

more than fully acclimatized at temperatures higher than 20.8 °C (Fig. 3b).

This study, for the first time, to our knowledge, has quantified the degree of the acclimatization of soil respiration to warming that varies with temperature itself. Other lines of evidence^{16,23} also support our result that respiratory acclimatization is stronger in the high rather than in the low temperature range. If this relationship holds over larger spatial and temporal scales, ecosystems in warm regions might acclimatize to global warming more than ecosystems in cold regions. Ecosystems in a future, warmer climate might acclimatize more than the current ecosystems do. If acclimatization is indeed caused by limited substrate supply, ecosystems with low soil carbon content may acclimatize to global warming more than ecosystems with a high carbon supply. That means ecosystems in cold regions with high carbon storage may be more closely coupled with climate than ecosystems in warm regions with low carbon storage, as suggested by other studies²⁰. In general, the spatial and temporal variability in respiratory responses to warming not only highlights the limitations of the Q_{10} approach to global scaling but also emphasizes the need for research on the biosphere–atmosphere interactions.

Interactions between climatic change and the carbon cycle are likely to be complicated by land-use change^{24,25}. In this study, the unwarmed, unclipped subplots probably represent undisturbed prairie under the current climate. The warmed, clipped subplots presumably represent a land-use scenario in a future, warmed climate. Soil temperature differed by 3.2 °C between the two treatments, suggesting that clipping or grazing exacerbates the effects of climatic warming. Despite the large temperature difference, observed soil respiration between the two scenarios was similar, owing to a decrease in Q_{10} values from 2.7 to 2.1. Thus, strong acclimatization of soil respiration to temperature may mitigate the

effects of land use and global warming on the terrestrial carbon cycle.

Overall, acclimatization of soil respiration to global warming counteracts the positive feedback between the carbon cycle and climatic warming and may act as a short-term mechanism of homeostasis in the earth system (Fig. 1). As carbon released via soil respiration accounts for approximately 10% of the carbon in the atmospheric pool²⁵, most modelling studies demonstrate a major feedback of soil respiration to reinforce global warming^{2–4,7}. On the basis of temperature sensitivity with a fixed Q_{10} value (for example, 2.0) across the globe, global warming by 2 °C is predicted to increase additional carbon release from soil by more than 10 Pg (10^{15} g) of C per year. The additional carbon release to the atmosphere results in more greenhouse effects, aggravating anthropogenic warming^{2,3}. Acclimatization of soil respiration to global warming has the potential to offset, at least partially, the additional carbon release that would be stimulated by a temperature increase; it thus weakens the climate to carbon cycle coupling. □

Methods

The experimental site is at the Central Redbed Plains²⁶. The grassland is dominated by C_4 grasses (*Schizachyrium scoparium*, *Sorghastrum nutans* and *Eragrostis* spp.) and C_3 forbs (*Ambrosia psilostachya* and *Xanthocephalum texanum*). Mean annual temperature is 16.3 °C with January being the coldest month (3.3 °C) and July the warmest (28.2 °C). Mean annual precipitation is 914 mm. The soil is part of the Nash–Lucien complex, which is characterized as having a low permeability rate, high available water capacity and a deep, moderately penetrable root zone²⁷.

The experiment used a paired, nested design with warming as a primary factor and clipping as a secondary factor. Five pairs of plots of 2 m × 2 m were established for warming and unwarmed control. Each plot was divided into four 1 m × 1 m subplots. Plants in two diagonal subplots were clipped at the height of 10 cm above the ground on 15 November 1999 and 28 July 2000 with the other two as unclipped control. An infrared radiator (Kalglo Electronics, Bethlehem, Pennsylvania) was used as the heating device, suspended 1.5 m above the ground. The device has been effectively used in other studies¹⁷.

Air temperature was monitored with sheltered thermocouples at the height of 25 cm in each plot. Soil temperature was measured at the depth of 2.5 cm in one clipped and one unclipped subplot in each plot. All the thermocouples were connected to a CR10 datalogger (Campbell Scientific, Logan, Utah). Air and soil temperatures were measured every 10 minutes. Averages within one hour were stored in SM196 Storage Module. Soil respiration was measured between 10:00 and 15:00 twice a month with an infrared gas analyser Li-Cor 6400 with an attachment of soil chamber Li-Cor 6400-09 (Li-Cor, Lincoln, Nebraska). We collected soil samples at the depth of 0–5 cm in one unclipped and one clipped subplot in each plot to measure soil moisture gravimetrically at the time of measuring respiration.

Received 30 April; accepted 16 August 2001.

- Houghton, J. T. et al. (eds) *Climate Change 2001: The Scientific Basis* 1–896 (Cambridge Univ. Press, Cambridge, 2001).
- Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A. & Totterdell, I. J. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature* **408**, 184–187 (2000).
- Lenton, T. M. Land and ocean carbon cycle feedback effects on global warming in a simple Earth system model. *Tellus B* **52**, 1159–1188 (2000).
- Woodwell, G. M. et al. Biotic feedbacks in the warming of the earth. *Clim. Change* **40**, 495–518 (1998).
- Sarmiento, J. Global change: That sinking feeling. *Nature* **408**, 155–156 (2000).
- Schimel, D. S. et al. Climate and nitrogen controls on the geography and timescales of terrestrial biogeochemical cycling. *Glob. Biogeochem. Cycles* **10**, 677–692 (1996).
- Jenkinson, D. S., Adams, D. E. & Wild, A. Model estimates of CO_2 emissions from soil in response to global warming. *Nature* **351**, 304–306 (1991).
- Kirschbaum, M. U. F. Will changes in soil organic matter act as a positive or negative feedback on global warming? *Biogeochemistry* **48**, 21–51 (2000).
- Lloyd, J. & Taylor, J. A. On the temperature dependence of soil respiration. *Funct. Ecol.* **8**, 315–323 (1994).
- Verville, J. H., Hobbie, S. E., Chapin, F. S. III & Hooper, D. U. Response of tundra CH_4 and CO_2 flux to manipulation of temperature and vegetation. *Biogeochemistry* **41**, 215–235 (1998).
- Liski, J., Ilvesniemi, H., Mäkelä, A. & Westman, C. J. CO_2 emissions from soil in response to climatic warming are overestimated—The decomposition of old soil organic matter is tolerant of temperature. *Ambio* **28**, 171–174 (1999).
- Lin, G., Ehleringer, J. R., Rygielwicz, P. T., Johnson, M. G. & Tingey, D. T. Elevated CO_2 and temperature impacts on different components of soil CO_2 efflux in Douglas-fir terraces. *Glob. Change Biol.* **5**, 157–168 (1999).
- Rustad, L. E. & Fernandez, I. J. Experimental soil warming effects on CO_2 and CH_4 flux from a low elevation spruce-fir forest soil in Maine, USA. *Glob. Change Biol.* **4**, 597–605 (1998).
- Peterjohn, W. T. et al. Responses of trace gas fluxes and N availability to experimentally elevated soil temperatures. *Ecol. Appl.* **4**, 617–625 (1994).
- Rustad, L. E. et al. A meta-analysis of the response of soil respiration, net nitrogen mineralization, and aboveground plant growth to experimental ecosystem warming. *Oecologia* **126**, 543–562 (2001).
- McHale, P. J., Mitchell, M. J. & Bowles, F. P. Soil warming in a northern hardwood forest: trace gas fluxes and leaf litter decomposition. *Can. J. Forest. Res.* **28**, 1365–1372 (1998).

17. Saleska, S. R., Harte, J. & Torn, M. S. The effect of experimental ecosystem warming on CO₂ fluxes in a montane meadow. *Glob. Change Biol.* **5**, 125–141 (1999).
18. Cheng, W., Zhang, Q., Coleman, D. C., Carroll, C. R. & Hoffman, C. A. Is available carbon limiting microbial respiration in the rhizosphere? *Soil Biol. Biochem.* **28**, 1283–1288 (1996).
19. Giardina, C. P. & Ryan, M. G. Evidence that decomposition rates of organic carbon in mineral soil do not vary with temperature. *Nature* **404**, 858–861 (2000).
20. Oechel, W. C. *et al.* Acclimation of ecosystem CO₂ exchange in the Alaskan Arctic in response to decadal climate warming. *Nature* **406**, 978–981 (2000).
21. Zogg, G. P. *et al.* Compositional and functional shifts in microbial communities due to soil warming. *Soil Sci. Soc. Am. J.* **61**, 475–481 (1997).
22. Atkin, O. K., Edwards, E. J. & Loveys, B. R. Response of root respiration to changes in temperature and its relevance to global warming. *New Phytol.* **147**, 141–154 (2000).
23. Kirshbaum, M. U. F. The temperature dependence of soil organic matter decomposition, and the effect of global warming on soil organic C storage. *Soil Biol. Biochem.* **27**, 753–760 (1995).
24. Striegl, R. G. & Wickland, K. P. Effects of a clear-cut harvest on soil respiration in a jack pine–lichen woodland. *Can. J. Forest. Res.* **28**, 534–539 (1998).
25. Raich, J. W. & Potter, C. S. Global patterns of carbon dioxide emissions from soils. *Glob. Biogeochem. Cycles* **9**, 23–36 (1995).
26. Tarr, J., Botkin, G., Rice, E. L., Carpenter, E. & Hart, M. A broad analysis of fifteen sites in the tall-grass prairie of Oklahoma. *Proc. Oklahoma Acad. Sci.* **60**, 39–42 (1980).
27. *National Cooperative Soil Survey* 1–151 (US Department of Agriculture, Soil Survey of McClain County, Oklahoma Agricultural Experiment Station, Stillwater, 1963).
28. Shaver, G. R. *et al.* Global warming and terrestrial ecosystems: a conceptual framework for analysis. *BioScience* **50**, 871–882 (2000).

Supplementary information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank S. Hu for his comments. This research was financially supported by the University of Oklahoma Research Council, the US National Science Foundation and the Department of Energy.

Correspondence and requests for materials should be addressed to Y.L. (e-mail: yluo@ou.edu).

The earliest known fully quadrupedal sirenian

D. P. Domning

Department of Anatomy, Howard University, Washington DC 20059, USA

Modern seacows (manatees and dugongs; Mammalia, Sirenia) are completely aquatic, with flipperlike forelimbs and no hindlimbs^{1,2}. Here I describe Eocene fossils from Jamaica that represent nearly the entire skeleton of a new genus and species of sirenian—the most primitive for which extensive postcranial remains are known. This animal was fully capable of locomotion on land, with four well-developed legs, a multivertebral sacrum, and a strong sacroiliac articulation that could support the weight of the body out of water as in land mammals. Aquatic adaptations show, however, that it probably spent most of its time in the water. Its intermediate form thus illustrates the evolutionary transition between terrestrial and aquatic life. Similar to contemporary primitive cetaceans³, it probably swam by spinal extension with simultaneous pelvic paddling, unlike later sirenians and cetaceans, which lost the hindlimbs and enlarged the tail to serve as the main propulsive organ. Together with fossils of later sirenians elsewhere in the world^{4–7}, these new specimens document one of the most marked examples of morphological evolution in the vertebrate fossil record.

Since 1990, abundant remains of sirenians, together with other marine taxa of early middle Eocene age (roughly 50 Myr ago), have been collected from Seven Rivers, Jamaica. The sediments (siltstones and sandstones) represent a lagoonal, estuarine or deltaic depositional environment. The age of the site was determined from molluscs (*Campanile*, *Eovasum*, *Paraseraphs*, *Velates*), and

corroborated by the presence of a primitive rhinoceros, *Hyrachyus* sp., of late early or early middle Eocene date⁸. The terrestrial fauna also includes an iguanian lizard⁹ and possibly a primate¹⁰. The sirenian fossils are found in five distinct bone-beds within a 5-m stratigraphic section referred to the Guys Hill Member of the Chapelton Formation, and occur as isolated bones and associated partial skeletons, with the remains of several individuals commingled. In the lower three bone-beds, all the sirenian remains appear to represent a single taxon (described here: Figs 1 and 2). These provide the first view of the overall anatomy and mode of locomotion of sirenians during their evolutionary transition from land-dwelling to obligatorily aquatic life.

Order Sirenia Illiger, 1811

Family Prorastomidae Cope, 1889

Pezosiren portelli, gen. et sp. nov.

Etymology. Generic name from Greek *pezos* (on foot, walking) and *Seiren* (Latin *Siren*, f., siren). Specific epithet in honour of Roger W. Portell, discoverer and co-investigator of the site.

Locality and horizon. Known only from Seven Rivers (about 15 km south of Montego Bay), parish of St James, western Jamaica; Guys Hill Member, Chapelton Formation, Yellow Limestone Group (early middle Eocene).

Holotype. US National Museum of Natural History, Paleobiology collection (USNM), no. 511925 (Fig. 2a): pair of partial mandibles from Seven Rivers bone-bed II.

Referred material. Specimens illustrated here (Fig. 2b–h) comprise sacrum (associated with 7 thoracic, 4 lumbar and 1 caudal vertebrae), USNM 517463, bone-bed II; left and right innominates, USNM 517464, bone-bed II; left femur with separated distal epiphysis, USNM 517465, bone-bed II; left tibia, USNM 517466, bone-bed III; intermediate phalanx, USNM 517467, bone-bed II. The following description also draws on several hundred other cranial and postcranial elements from Seven Rivers bone-beds II–IV, not formally referred at this time, that remain to be catalogued in the USNM, University of the West Indies, and Florida Museum of Natural History collections.

Diagnosis. The diagnosis for the genus is the same as for the species until other species are described. *Pezosiren portelli* differs from *Prorastomus sirenoides*, the only other named member of the family, in having (1) a sagittal crest; (2) an auditory meatus that is about as wide anteroposteriorly as high; (3) a periotic that is not fused to the alisphenoid; (4) an unenlarged P₁; and (5) a horizontal mandibular ramus whose ventral border is turned down anteriorly. Character states 1 and 4 are primitive within the Sirenia; character states 2, 3 and 5 are considered derived¹¹.

Description. *Pezosiren* was a pig-sized quadruped (estimated total length 2.1 m) with a relatively short (but not extremely compressed) neck, a long, barrel-shaped trunk, short legs and a substantial (but not powerfully muscled) tail (Fig. 1). The skull (estimated condylobasal length 26.5 cm, zygomatic width 14.5 cm) is of clearly sirenian form, with a prominent although undeflected rostrum, large mesorostral fossa, large nasals, horizontally jutting supraorbital processes, and laterally projecting zygomatic arches. A weak sagittal crest is present; this appears to be absent in the holotype of *Prorastomus sirenoides* but may have been removed by erosion or preparation. The sphenopalatine region is primitively long, and lacks the enlarged, stout, ventrally projecting pterygoid processes characteristic of later sirenians. An alisphenoid canal is present, and the foramen ovale seems to be completely enclosed by bone.

The mandible (Fig. 2a) is long (19 cm) and slender, with its ventral border strongly turned down anteriorly; the symphysis is long, deep and mediolaterally compressed; the incisors and canines are arranged in parallel, parasagittal rows; several mental foramina are present; and the mandibular foramen is small, not exposing the dental capsule. The dental formula is uncertain (the toothrows are incomplete anteriorly) but was presumably 3.1.5.3 as in all other

Eocene sirenians¹². The first upper incisor is enlarged to form a small tusk with subconical enamel crown. The alveolus interpreted as that of the lower canine is large, but not demonstrably divided for a two-rooted tooth as in *Prorastomus*. The first lower premolar alveolus is single and not enlarged, further contrasting with the large, two-rooted first premolar of *Prorastomus*. The tooth in front of M₁ is single-rooted, indicating replacement at the last (fifth) premolar locus.

The vertebral column (judging in part from two incomplete but

overlapping series of associated trunk vertebrae) evidently comprised 7 cervical, 20 thoracic, 4 lumbar, 4 sacral, and probably a couple of dozen caudal vertebrae. The atlas has not been recovered. The axis and other cervical vertebrae grossly resemble those of Palaeocene condylarths such as *Ectoconus*¹³, rather than being as compressed anteroposteriorly as in later sirenians. Anterior thoracics have tall neural spines, indicating the presence of a nuchal ligament that was able to support the weight of the head and neck on land. The thoracic, lumbar and sacral vertebrae are remarkable for

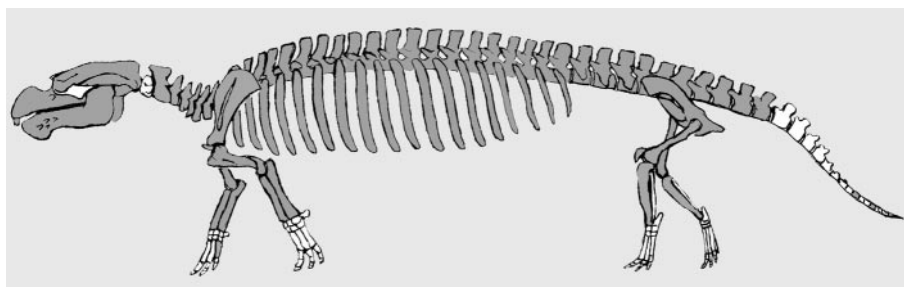


Figure 1 Reconstructed composite skeleton of *Pezosiren portelli*. Lateral view, length roughly 2.1 m. Shaded elements are represented by fossils; unshaded elements (jugal,

atlas, fibula, most bones of the feet, and most caudal vertebrae) are not. The length of the tail, and the form and posture of the feet are partly conjectural.



Figure 2 *Pezosiren portelli*. **a**, Partial right mandible (holotype, USNM 511925); medial view, showing location of small mandibular foramen (white arrow). Approximate outlines of restored broken portions are shown. **b**, Dorsal view of sacrum (USNM 517463); pleurapophyses are missing on S3–S4. **c**, Right innominate (USNM 517464); ventromedial view. **d**, Right and left innominates (USNM 517464); ventral view. The

apparently ventral direction of the acetabula is an artefact of dorsoventral crushing. **e**, Subadult left femur (USNM 517465); anterior view. **f**, Left tibia (USNM 517466); anterior view. **g**, Left tibia (USNM 517466); lateral view. Note oblique, distolaterad-facing distal articular surface of tibia. **h**, Intermediate phalanx (USNM 517467); dorsal view. Scale bars, 5 cm (**a–g**); 1 cm (**h**).

having large, horizontally projecting flanges at the tips of the neural spines and prominent mammillary processes.

The transverse processes (pleurapophyses) of the (usually unfused) sacral vertebrae (Fig. 2b) are connected by rigid lateral articulations as in the cetacean *Rodhocetus*¹⁴. In older individuals the sacra may be at least partly fused. The anterior sacra bear large articular surfaces for the ilium, as in land mammals. Caudal vertebrae lack the enlarged transverse processes of later sirenians; the small transverse processes of the more anterior ones were pierced by foramina or notches, as in several other aquatic mammals but not in later sirenians. The ribs are swollen (pachyostotic) and composed wholly of dense bone (osteosclerotic), as in other sirenians; their shafts are subcylindrical and lack costal grooves. The sternum comprises at least three broad, dorsoventrally flattened elements, in contrast to the more primitive sternbrae seen in the otherwise more derived sirenian *Protosiren smithae*³.

The scapular blade is broad, not sickle-shaped as in protosirenids and primitive dugongids. There is no clavicle. The humerus is remarkably broad distally. The ulnar shaft is flattened, and was evidently connected immovably (although not fused) to the radius. The radius apparently lay anteromedial to the ulna, but is not well preserved; its exact form and relationships are unclear, as are the structure and orientation of the wrist joint. The pelvis (Fig. 2c, d) is long and narrow (total length 243 mm), resembling those of primitive land mammals, with a moderately broad ilium and an elongate ischial tuberosity. The acetabulum, obturator foramen, and pubic symphysis are well developed. The femur (Fig. 2e; length 166 mm) is robust, with a distinct third trochanter.

The patella is large and shaped like a teardrop. The tibia (length 130 mm; Fig. 2f, g) greatly resembles those of *Ectoconus*¹³ or other primitive ungulates, with a long and prominent cnemial crest, medial torsion of the shaft relative to the proximal end, and a distal articular surface that faces obliquely distolaterad¹⁵. Of the feet, only one metapodial, one proximal and one intermediate phalanx have been recovered; whether they represent the front or back feet is unknown. These bones are short and flattened like those of terrestrial ungulates, and show no signs of paddle-like elongation (Fig. 2h). The skeleton is shown here conjecturally with a digitigrade foot posture consistent with the presumed descent of sirenians from semicursorial "condylarths"^{16,17}; however, a secondarily plantigrade posture such as that of otters (see 'Functional interpretation' below) cannot be ruled out.

Affinities. *Pezosiren portelli* is clearly a sirenian on the basis of its lack of paranasal air sinuses and its possession of a prominent rostrum, retracted and enlarged external nares, premaxilla-frontal contact, longitudinally aligned incisor arcades and extensive pachyosteosclerosis in the skeleton (see ref. 11 for relevant phylogenetic analysis). It is placed in the basal (and paraphyletic) sirenian family, the Prorastomidae, on the basis of its primitively undeflected rostrum, unenlarged pterygoid process and unexposed mandibular dental capsule. This hitherto monotypic family was erected for *Prorastomus sirenioides*, which was founded on a single skull, mandible and atlas vertebra from the parish of Trelawny, Jamaica, about 40 km east-southeast of Seven Rivers, in the earliest middle Eocene Stettin Member of the Chapelton Formation (a slightly earlier horizon than that of *Pezosiren*)¹⁸. Hardly any other remains (and no appendicular bones) referable to this family were discovered before the Seven Rivers finds. *Prorastomus* differs from *Pezosiren* most obviously in that the ventral border of the mandibular ramus is not downturned anteriorly (a primitive condition). Although *Pezosiren* seems to share some synapomorphies with more derived sirenians to the exclusion of *Prorastomus*, other character states are noncongruent and/or poorly understood, so a definitive conclusion on its phylogenetic position would be premature.

Functional interpretation. The tall anterior thoracic neural spines, multivertebral sacrum, strong sacroiliac articulation and robust appendicular elements with fully developed joint surfaces are

exactly comparable to those of similar-sized land mammals, and show clearly that *Pezosiren* was capable of supporting its body weight out of water. It was a true sirenian, however, and its clear aquatic specializations (retracted nasal opening, lack of paranasal air sinuses and thorax with numerous heavy, pachyosteosclerotic ribs for ballast^{19,20}) suggest that it spent most of its time in the water, probably feeding as well as resting there.

Unlike modern sirenians and cetaceans, it did not propel itself in the water by means of up-and-down undulations of a powerful tail. Its tail vertebrae lacked the broad transverse processes needed for large tail muscles, and the tail was probably otter-like, with some dorsoventral flattening but no distinct caudal fin or flukes. Main propulsion was apparently provided instead by forceful extension of the spine, and backward and upward thrusts of the hind feet together—in other words, simultaneous pelvic paddling combined with spinal undulation²¹. This mode of swimming is typical of otters and is also thought to have occurred in primitive whales such as *Ambulocetus*^{3,22}. Strong spinal extension of this sort might explain the peculiar flanges on the tips of the thoracic, lumbar and sacral neural spines, as well as the large mammillary processes; powerful longissimus muscles would have inserted on both sites. □

Received 2 February; accepted 16 August 2001.

- Domning, D. P. The readaptation of Eocene sirenians to life in water. *Hist. Biol.* **14**, 115–119 (2000).
- Domning, D. P. In *Secondary Adaptation of Tetrapods to Life in Water* (eds Mazin, J.-M. & de Buffrénil, V.) 151–168 (Dr Friedrich Pfeil, Munich, 2001).
- Thewissen, J. G. M., Hussain, S. T. & Arif, M. Fossil evidence for the origin of aquatic locomotion in archaeocete whales. *Science* **263**, 210–212 (1994).
- Domning, D. P. & Gingerich, P. D. *Protosiren smithae*, new species (Mammalia, Sirenia), from the late Middle Eocene of Wadi Hitan, Egypt. *Contr. Mus. Paleontol. Univ. Michigan* **29**, 69–87 (1994).
- Siegfried, P. Das Femur von *Eotheroides libyca* (Owen) (Sirenia). *Palaeontol. Zs.* **41**, 165–172 (1967).
- Lepsius, G. R. *Halitherium schinzi*, die fossile Sirene des Mainzer Beckens. *Abh. Mittelrheinischen Geol. Ver.* **1**, 1–200 (1882).
- Abel, O. Die Sirenen der mediterranen Tertiärbildungen Österreichs. *Abh. Geol. Reichsanst. Wien* **19**, 1–223 (1904).
- Domning, D. P., Emry, R. J., Portell, R. W., Donovan, S. K. & Schindler, K. S. Oldest West Indian land mammal: rhinocerotoid ungulate from the Eocene of Jamaica. *J. Vertebr. Paleontol.* **17**, 638–641 (1997).
- Pregill, G. K. Eocene lizard from Jamaica. *Herpetologica* **55**, 157–161 (1999).
- MacPhee, R. D. E., Flemming, C., Domning, D. P., Portell, R. W., & Beatty, B. Eocene primate petrosal from Jamaica: morphology and biogeographical implications. *J. Vertebr. Paleontol.* **19** (Suppl.), 61A (1999).
- Domning, D. P. A phylogenetic analysis of the Sirenia. *Proc. San Diego Soc. Nat. Hist.* **29**, 177–189 (1994).
- Domning, D. P., Morgan, G. S., & Ray, C. E. North American Eocene sea cows (Mammalia: Sirenia). *Smithson. Contr. Paleobiol.* **52**, 1–69 (1982).
- Matthew, W. D. Paleocene faunas of the San Juan Basin, New Mexico. *Trans. Am. Phil. Soc.* **30**, 1–510 (1937).
- Gingerich, P. D., Raza, S. M., Arif, M., Anwar, M. & Zhou, X. New whale from the Eocene of Pakistan and the origin of cetacean swimming. *Nature* **368**, 844–847 (1994).
- Domning, D. P. Supposedly unique features of the desmostylians tibia are also found in other mammals. *Bull. Ashoro Mus. Paleontol.* **2**, 39–42 (2001).
- Thewissen, J. G. M. & Domning, D. P. The role of phenacodontids in the origin of the modern orders of ungulate mammals. *J. Vertebr. Paleontol.* **12**, 494–504 (1992).
- Liu, F. R. & Miyamoto, M. M. Phylogenetic assessment of molecular and morphological data for eutherian mammals. *Syst. Biol.* **48**, 54–64 (1999).
- Savage, R. J. G., Domning, D. P. & Thewissen, J. G. M. Fossil Sirenia of the West Atlantic and Caribbean region. V. The most primitive known sirenian, *Prorastomus sirenioides* Owen, 1855. *J. Vertebr. Paleontol.* **14**, 427–449 (1994).
- Domning, D. P. & de Buffrénil, V. Hydrostasis in the Sirenia: quantitative data and functional interpretations. *Mar. Mamm. Sci.* **7**, 331–368 (1991).
- Taylor, M. A. Functional significance of bone ballast in the evolution of buoyancy control strategies by aquatic tetrapods. *Hist. Biol.* **14**, 15–331 (2000).
- Fish, F. E. Transitions from drag-based to lift-based propulsion in mammalian swimming. *Am. Zool.* **36**, 628–641 (1996).
- Thewissen, J. G. M., Madar, S. I. & Hussain, S. T. *Ambulocetus natans*, an Eocene cetacean (Mammalia) from Pakistan. *Cour. Forsch. Inst. Senckenberg* **191**, 1–86 (1996).

Acknowledgements

I thank J. Bailey, B. Beatty, D. Da Silva, H. Dixon, S. K. Donovan, R. J. Emry, C. Flemming, F. Grady, H. and J. Halvorson, J. Herrera, K. Hickey-Commins, S. Hutchens, S. Jabo, D. Jones, I. A. Koretsky, J. Kramer, B. J. MacFadden, C. MacGillivray, R. D. E. MacPhee, S. Mitchell, R. W. Portell, T. Radenbaugh, K. S. Schindler, T. A. Stemmann, C. Terranova, and B. R. and J. Toomey. Field work was funded by the National Geographic Society; B. and R. Toomey; the Potomac Museum Group; the University of the West Indies; the American Museum of Natural History; and R. Liberman.

Correspondence and requests for materials should be addressed to the author (e-mail: ddomning@fac.howard.edu).

A general model for ontogenetic growth

Geoffrey B. West*†, James H. Brown†‡ & Brian J. Enquist§

* Theoretical Division, MS B285, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

† The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

‡ Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA

§ Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

Several equations have been proposed to describe ontogenetic growth trajectories for organisms justified primarily on the goodness of fit rather than on any biological mechanism^{1–6}. Here, we derive a general quantitative model based on fundamental principles^{7–9} for the allocation of metabolic energy between maintenance of existing tissue and the production of new biomass. We thus predict the parameters governing growth curves from basic cellular properties¹⁰ and derive a single parameterless universal curve that describes the growth of many diverse species. The model provides the basis for deriving allometric relationships for growth rates and the timing of life history events^{2,11,12}.

Ontogenetic development is fuelled by metabolism and occurs primarily by cell division. Incoming energy and materials from the environment are transported through hierarchical branching network systems to supply all cells. These resources are transformed into metabolic energy, which is used for life-sustaining activities. During growth, some fraction of this energy is allocated to the production of new tissue. Thus, the rate of energy transformation is the sum of two terms, one of which represents the maintenance of existing tissue, and the other, the creation of new tissue. This is expressed by the conservation of energy equation:

$$B = \sum_c \left[N_c B_c + E_c \frac{dN_c}{dt} \right] \quad (1)$$

The incoming rate of energy flow, B , is the average resting metabolic rate of the whole organism at time t , B_c is the metabolic rate of a single cell, E_c is the metabolic energy required to create a cell and N_c is the total number of cells; the sum is over all types of tissue. Possible differences between tissues are ignored and some average typical cell is taken as the fundamental unit. The first term, $N_c B_c$, is the power needed to sustain the organism in all of its activities, whereas the second is the power allocated to production of new cells and therefore to growth. E_c , B_c , and the mass of a cell, m_c , are assumed to be independent of m remaining constant throughout growth and development.

At any time t the total body mass $m = m_c N_c$, so equation (1) can be written as

$$\frac{dm}{dt} = \left(\frac{m_c}{E_c} \right) B - \left(\frac{B_c}{E_c} \right) m \quad (2)$$

Now, if $B = B_0 m^{3/4}$, where B_0 is constant for a given taxon, then

$$\frac{dm}{dt} = am^{3/4} - bm \quad (3)$$

with $a \equiv B_0 m_c / E_c$ and $b \equiv B_c / E_c$. The 3/4 exponent is well supported by data on mammals^{1,13}, birds¹⁴, fish^{15,16}, molluscs¹⁷ and plants¹⁸. Although some mammals may show fluctuations around 3/4-power scaling owing to 'growth spurts' (ref. 1), the 3/4 exponent describes the overall allometry of B from birth to reproductive maturity. For altricial birds, hatchlings are supplied with a store of metabolically inert water which is expended during growth, and when this is taken

into account in relating N_c to m , $B \propto m^{3/4}$ (ref. 14).

Recently, a model was developed for understanding the 3/4 exponent and, more generally, the ubiquitous 1/4 power occurring in biological allometry^{7,8}. It is based on the premise that the tendency of natural selection to optimize energy transport has led to the evolution of fractal-like distribution networks. The 3/4 exponent was shown to be related to the scaling of the total number (N_t) of terminal units (capillaries) in the network: $B \propto N_t \propto m^{3/4}$. In contrast, the total number of cells, $N_c \propto m$. Thus, the reason for the different exponents of m in the two terms on the right-hand side of equation (3) is that the network constrains the total number of supply units (capillaries) to scale differently from the total number of cells supplied^{7,8}. This imbalance between supply and demand ultimately limits growth. If the exponents were the same, then $dm/dt \neq 0$ and organisms would continue to grow indefinitely. We therefore have a fundamental explanation for the origin of determinate growth in which an asymptotic maximum body size (M) is reached. This occurs when $dm/dt = 0$, giving $M = (a/b)^4 = (B_0 m_c / B_c)^4$. Thus, the variation in M among species within a taxon, where B_0 and m_c do not change, is determined by the systematic variation of the *in vivo* cellular metabolic rate, B_c , which scales as $M^{-1/4}$. Within a taxon B_0 , m_c and E_c are approximately constant, so a should be approximately independent of M , whereas $b (= a/M^{1/4})$ should scale as $M^{-1/4}$. Between groups, however, a should vary, principally reflecting variations in B_0 . Equation (3) can therefore be re-expressed as

$$\frac{dm}{dt} = am^{3/4} \left[1 - \left(\frac{m}{M} \right)^{1/4} \right] \quad (4)$$

Although equations (3) and (4) are superficially similar in structure to that of von Bertalanffy⁶, they differ significantly in that they are derived from basic principles so that the parameters governing growth, a and b , are directly calculable from fundamental cellular parameters.

A classical sigmoidal curve (see Supplementary Information) is obtained from integrating equation (4):

$$\left(\frac{m}{M} \right)^{1/4} = 1 - \left[1 - \left(\frac{m_0}{M} \right)^{1/4} \right] e^{-at/4M^{1/4}} \quad (5)$$

Here, m_0 is the mass at birth ($t = 0$). In Fig. 1 we plot some examples of m versus t for four very different animals and fit the data using equation (5). Values of a , m_0 and M for these and several other species can be found in Table 1. Consistent with our predictions, a varies only modestly within a taxon, whereas across taxa, $a \propto B_0$, as is confirmed by a positive correlation with B_0 (coefficient of correlation, $r^2 = 0.82$; $n = 5$; $P = 0.035$)¹¹. Perhaps more significantly, the magnitudes of a and b can be independently determined from fundamental parameters of the cell. The energy content of mammalian tissue has been measured to be about $7 \times 10^6 \text{ J Kg}^{-1}$ (refs 11, 19), so, if $m_c \approx 3 \times 10^{-9} \text{ g}$, the energy needed to create a cell

Table 1 Values of several parameters for various organisms

Organism	a	m_0	M	Slope
Cow	0.28	33.3 kg	442 kg	1.08
Pig	0.31	0.90 kg	320 kg	1.08
Rabbit	0.36	0.12 kg	1.35 kg	1.34
Guinea pig	0.21	5 g	840 g	0.91
Rat	0.23	8 g	280 g	1.07
Shrew	0.83	0.3 g	4.2 g	0.98
Heron	1.56	3 g	2.7 kg	1.04
Hen	0.47	43 g	2.1 kg	0.72
Robin	1.9	1 g	22 g	1.03
Cod	0.017	0.1 g	25 kg	1.01
Salmon	0.026	0.01 g	2.4 kg	1.01
Guppy	0.10	0.008 g	0.15 g	1.04
Shrimp	0.027	0.0008 g	0.075 g	0.82

a , see equation (3); m_0 , birth mass; M , asymptotic mass. Also shown are the negative mean values of the slopes of plots of $\ln[R(t)/R(0)]$ versus $at/4M^{1/4}$, which is predicted to have a universal value of 1; $R = [1 - (m/M)^{1/4}]$ is the proportion of metabolic power devoted to growth.

in vivo, $E_c \approx 2.1 \times 10^{-5}$ J. Taking $B_0 \approx 1.9 \times 10^{-2}$ W (ref. 11) then gives $a \equiv B_0 m_c / E_c \approx 0.25$ g^{1/4} per day, in good agreement with data (Table 1). Asymptotic masses, M , or equivalently b , could be predicted if B_c *in vivo* were independently known.

Equation (5) suggests powerful ways of plotting the data that reveal universal properties of growth. If the dimensionless mass ratio, $r \equiv (m/M)^{1/4}$, is plotted against a dimensionless time variable, $\tau \equiv at/4M^{1/4} - \ln[1 - (m_0/M)^{1/4}]$, then equation (5) predicts that all species, regardless of taxon, cellular metabolic rate (B_c), or mature body size (M), should fall on the same parameterless universal curve $r = 1 - e^{-\tau}$. Data for a wide variety of animals (mammals, birds, fish, crustacea) clearly show such universal properties, as illustrated by Fig. 2. We note that τ includes an adjustment for m being non-zero at birth ($t = 0$). An alternative way of exhibiting this universality is to introduce $R \equiv [1 - (m/M)^{1/4}] = 1 - r$, in which case equation (5) becomes $R(t) = R(0)e^{-at/4M^{1/4}}$. A plot of $\ln[R(t)/R(0)]$ versus $at/4M^{1/4}$ should yield a universal straight line of slope -1 . The slopes for 13 species of animals are all very close to -1 , their mean being -0.99 ± 0.04 ; see Table 1.

The quantities, r and R , have an elegant interpretation as the relative proportions of total available metabolic power (B) that, respectively, fuel maintenance and growth. To see this, note that from equation (1), relative maintenance is given by $N_c B_c / B = (B_c / B_0 m_c) m^{1/4} = (b/a) m^{1/4} = (m/M)^{1/4} = r$, independently of any other parameter. Thus, Fig. 2 represents the proportion of power devoted to maintenance and other activities plotted as a function of dimensionless time, τ . Similarly, the proportion of total metabolic power devoted to growth is $1 - r = [1 - (m/M)^{1/4}] = R$. This therefore has a universal exponentially decreasing behaviour as a function of $at/4M^{1/4}$ throughout

ontogeny. Consequently, the proportion of B used for growth is the same for all species at the same stage of development, as measured relative to their asymptotic mass, M . For example, for all organisms, $R \approx 50\%$ ($R = 1 - (1/15)^{1/4} \approx 0.49$) when $m = M/15$, whereas it is around 16%, when $m = M/2$. These occur at very different temporal stages of ontogenetic development for different species. For example, at birth, a cow weighing about 40 kg ($\sim M/15$ if $M \approx 600$ kg; ref. 1), is expending about half of its metabolic power growing. After one year it has reached about half of its adult size ($m = M/2 \approx 300$ kg) and is allocating only 16% to growth. A codfish, however, almost always dies well before reaching $M/2$ ($M \approx 25$ kg; ref. 20); its maximum size is typically 6 kg ($\sim M/4$) reached at an age of around 10 years. Furthermore, it is already about 6 years old when its mass is only $M/15$ (~ 1.5 kg), at which stage it is still using about half of its metabolic power in continuing to grow.

For an organism with indeterminate growth, a substantial percentage of total metabolic energy is allocated to growth when its energy is integrated over its total lifespan. The lifetime allocation to growth is $E_g = \int E_c (dN_c/dt) dt = (E_c/m_c) \int (dm/dt) dt = (E_c/m_c)(m_m - m_0)$, where m_m is the maximum mass reached by the organism. As $m_m \gg m_0$ and $E_c/m_c = B_0/a$, $E_g \approx (B_0 M/a) r_m^4$, where $r_m = (m_m/M)^{1/4}$. The total metabolic energy used in a lifetime is $E_{tot} = \int B dt \approx -4(B_0 M/a)[r_m^3/3 + r_m^2/2 + r_m + \ln(1 - r_m)]$. For cod this gives $E_g/E_{tot} \approx 0.41$, so 41% of its total lifetime energy is spent on growth. For an organism with determinate growth, $r_m \approx 1$, so $E_g \approx (B_0 M/a)$. To calculate their E_{tot} , we first integrate B up to a time, t_m , corresponding to $m = (1 - \epsilon)M$ with $\epsilon \ll 1$; typically ϵ is chosen to be 0.01 or 0.05. This gives a contribution $(4B_0 M/a)[\ln(4/\epsilon) - 11/6]$. To this is added the energy used from t_m until death at $t = t_d$. Neglecting growth during this period gives

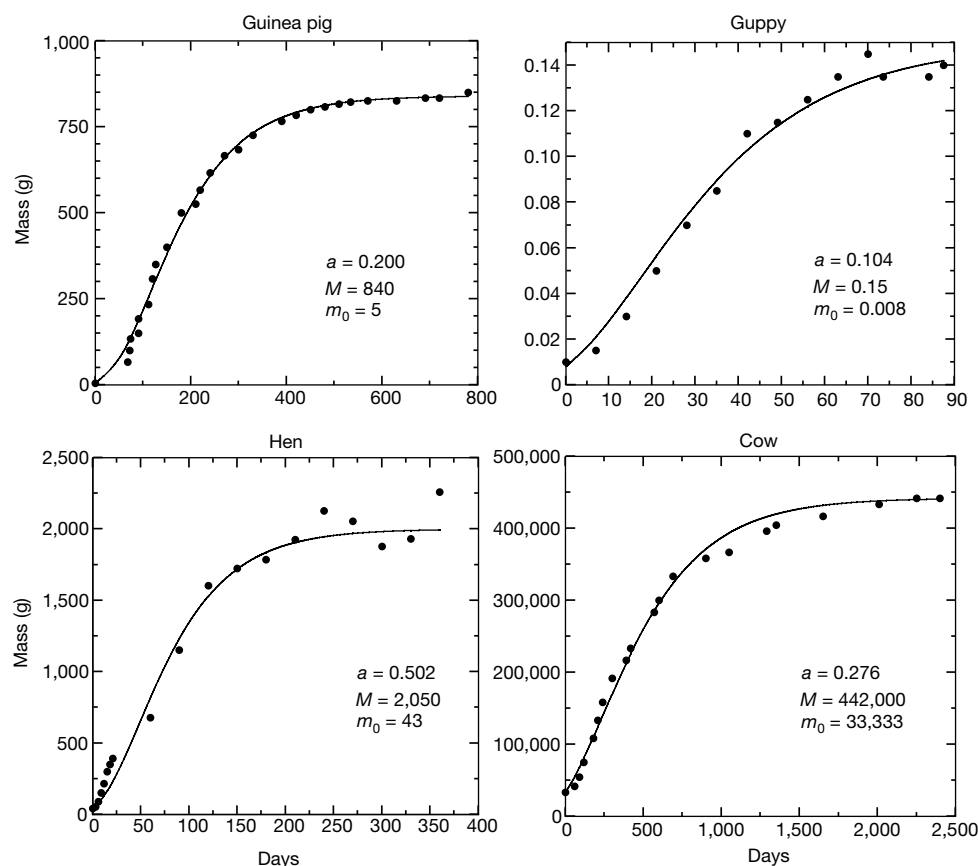


Figure 1 Four typical examples of fits to growth curves (solid lines) using equation (5). For definition of growth parameter a , see equation (3). M , asymptotic mass; m_0 , birth mass.

$(B_0 M/a)b(t_d - t_m)$. Typically, $b(t_d - t_m) \gg 4[\ln(4/\epsilon) - 11/6]$ so $E_g/E_{\text{tot}} \approx b(t_d - t_m) \ll 1$. For example, for a cow that lives for 20 years, we obtain $E_g/E_{\text{tot}} \approx 1\%$. On the other hand, growth accounts for almost 10% of its total metabolic energy expended before maturity.

R is maximal at birth, where $R(0) = 1 - (m_0/M)^{1/4}$. If $m_0/M < 1/16$, then $R(0) > 1/2$, so, for the vast majority of organisms, more than half of their metabolic power at birth is used for growth. For fish with external fertilization, and $m_0 \ll M$, $R(0) \approx 1$, so nearly all metabolism at hatching is used for growth. The point of inflection, where growth rate is maximal ($d^2m/dt^2 = 0$), occurs when $m = (3/4)^4 M \approx (1/3)M$ at which point $dm/dt = (27/256)aM^{3/4}$ and $R = 1/4$, independent of M . For some indeterminate growers this is never reached and growth continues to accelerate throughout life. On the other hand, for a 1 kg determinate growing mammal this gives $dm/dt \approx 6$ g per day, in good agreement with data¹.

Three important points need clarification. The first is cell replacement. Throughout ontogeny cells are dying and being replaced by mitosis. The power required for this is included in B_c , the average metabolic rate of a cell *in vivo*. If dN_c^d/dt is the cellular mortality rate, then the power required to replace cells that have died is $E_c dN_c^d/dt$. The cell death rate is proportional to the number of cells present, so $dN_c^d/dt = \gamma N_c$, where γ is the inverse lifetime of a typical cell. Thus, an amount γE_c must be contributed from B_c to replace the average cell; $b = B_c/E_c$, so γ represents that portion of b attributable to cell replacement. For a 50 kg mammal, a typical cell lives for around 2 months¹⁰, although there is considerable variation among tissue types; thus, $\gamma \approx 0.02 \text{ days}^{-1}$. Taking $E_c \approx 2 \times 10^{-5} \text{ J}$ gives $\gamma E_c \approx 4 \times 10^{-12} \text{ W}$, which is comparable to $B_c = B_0 m_c/M^{1/4}$. Thus, $\gamma \approx B_c/E_c = b = aM^{-1/4}$ so, across species, the lifetime of cells

increases as $M^{1/4}$. Moreover, once growth ceases, a substantial proportion of metabolism is devoted to cell replacement.

The second point is the difference between determinate and indeterminate growth. The two cases are distinguished by whether the organism reaches t_m and approaches an asymptotic size, M , before death (see Supplementary Information). From equation (5), $t_m \approx (4M^{1/4}/a) \ln[4\{1 - (m_0/M)^{1/4}\}/\epsilon]$. Except for a small logarithmic modulation, $t_m \propto M^{1/4}$, in agreement with data^{2,11}. A related time is the doubling time, T , which is the time taken to increase from m to $2m$. This is given by $T = (4a/M^{1/4}) \ln[\{1 - (m/M)^{1/4}\}/\{1 - (2m/M)^{1/4}\}]$, which, when $(m/M)^{1/4} \ll 1$, reduces to $T \approx (4a)(2^{1/4} - 1)m^{1/4}$, scaling as empirically observed^{4,5}. Organisms with determinate growth typically reproduce only after attaining nearly asymptotic size ($m \approx M$, when $t > t_m$) (ref. 2). For them, t_m can be equated with the age to first reproduction and there is no need during growth to modify the above equations to reflect energy allocation to reproduction. After maturation, reproduction is assumed to be fuelled by metabolic scope where metabolic rate is increased severalfold above the resting level to fuel activities such as thermoregulation and migration, as well as reproduction¹¹.

The third point concerns energy allocation to reproduction in indeterminate growth. Once such organisms reach their age for first reproduction (t_r), a significant fraction of metabolic rate is devoted to reproduction, and the growth rate is reduced. Consider the case of oviparous organisms^{2,3}. Egg production can be incorporated into equation (1) by adding a term $E_e dN_e/dt$ to the right-hand side when $t \geq t_r$; E_e is the energy needed to create a single egg cell and N_e the number created by time t . Data from oviparous ectothermic invertebrates and vertebrates^{11,21} support the assumption that, during a spawning period $\Delta t = t_s$, the mass of a clutch,

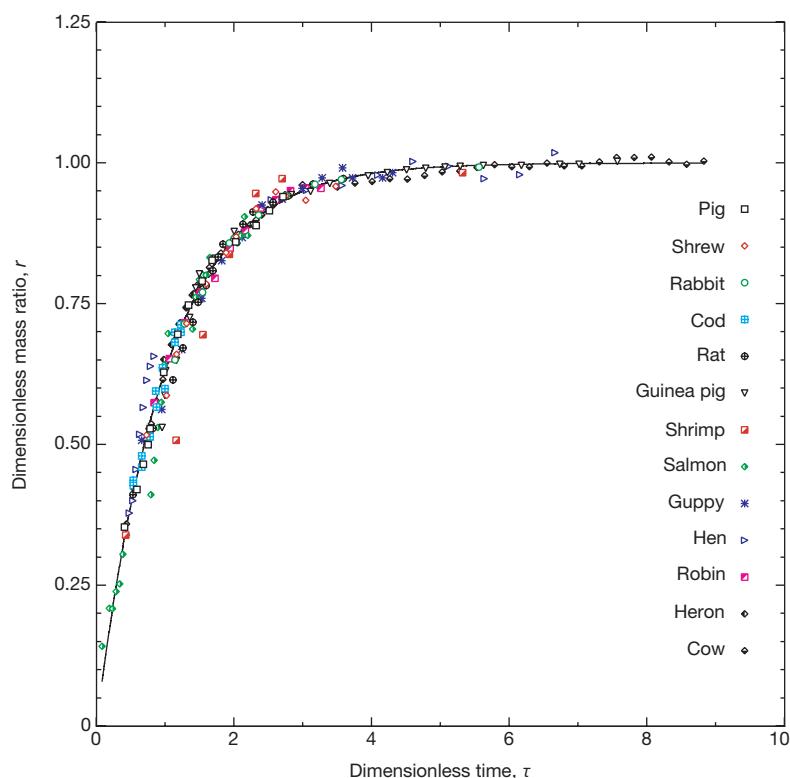


Figure 2 Universal growth curve. A plot of the dimensionless mass ratio, $r = 1 - R \equiv (m/M)^{1/4}$, versus the dimensionless time variable, $\tau = (at/4M)^{1/4} - \ln[1 - (m_0/M)^{1/4}]$, for a wide variety of determinate and indeterminate species. When plotted in this way, our model predicts that growth curves for all organisms should fall on

the same universal parameterless curve $1 - e^{-\tau}$ (shown as a solid line). The model identifies r as the proportion of total lifetime metabolic power used for maintenance and other activities.

$m_k = m_c \Delta N_c$, is a constant fraction, λ , of body mass: $m_k \approx \lambda m$. Thus $E_c dN_c/dt \approx E_c \Delta N_c/\Delta t \approx \lambda E_c m/m_c t_s$. The energy density of egg and other cells are similar, $E_c/m_c \approx E_c/m_c$, so this additional term becomes $E_c dN_c/dt \approx (\lambda E_c/t_s) N_c$. This is proportional to N_c and so has the identical structure to the maintenance term, $N_c B_c$, in equation (1). Thus, the solution, equation (5), is the same after reproduction ($t > t_r$) as before ($t < t_r$), except that B_c is replaced by $(B_c + \lambda E_c/t_s)$. In subsequent equations, a therefore remains the same before and after t_r , whereas b changes to $b' \equiv (b + \lambda/t_s)$. Consequently, because egg production continues throughout life, the actual asymptotic mass decreases from $M = (a/b)^4$ to $M' = (a/b')^4 = (1 + \lambda/bt_s)^{-4} M$. As an example, our fit to cod data ($M' \approx 25$ kg) gives $b' \approx 1.3 \times 10^{-3} \text{ days}^{-1}$. To get a rough estimate for the reproductive contribution we take $\lambda \approx 10\%$ (refs 11, 21) and $t_s \approx 100$ days and obtain $\lambda/t_s \approx 1 \times 10^{-3} \text{ days}^{-1}$, giving $b \approx 0.3 \times 10^{-3} \text{ days}^{-1}$. This value indicates that reproduction represents a significant portion of energy allocation. Thus, the proportion of maintenance energy allocated to reproduction relative to other activities, $E_c(dN_c/dt)/N_c B_c \approx \lambda/bt_s$, could be as much as a factor of 3. Consequently $M'/M = (1 + \lambda/bt_s)^{-4}$ could be as small as 10^{-2} . Thus $M \gg m$, so, for times before first reproduction ($t < t_r$), the solution is insensitive to $b = a/M^{1/4}$ and growth is determined primarily by a . In general, separate equations operate before and after t_r ; for most indeterminate growers, however, t_r is much smaller than lifespan, $t_r \ll t_d$, so growth is well approximated by a single equation—equations (4) or (5)—for all t but with b' (and M') replacing b (and M). These equations therefore apply to indeterminate and determinate growers with maintenance including reproduction and M being interpreted as M' in Table 1 and Fig. 2.

We have derived a very general growth equation from first principles on the basis of the conservation of metabolic energy, the allometric scaling of metabolic rate, and the energetic cost of producing and maintaining biomass (cells). The framework differs from recent work that has focused more on trade-offs involving reproduction and mortality^{2–4,22,23}. Our model attributes the slowing of growth as body size increases to limitations on the capacity of networks to supply sufficient resources to support further increase in body mass. Its power is demonstrated by its ability to quantitatively predict growth curves for both determinate and indeterminate growers, oviparous and viviparous species, ectotherms and endotherms, vertebrates and invertebrates (Fig. 2). The model can be extended to plants. Previously²⁴, a simpler version was presented for trees that included only the second term in equation (1). This was adequate largely because the first term only becomes important at large body sizes and only a small proportion of the trees had masses that approached the asymptotic value. Perhaps the most appealing and powerful feature of our model is that the parameters of the growth equation can be derived from fundamental cellular properties and predicted quantitatively from metabolic measurements that are not directly related to growth. □

Received 2 January; accepted 21 August 2001.

1. Brody, S. *Bioenergetics and Growth* (Hafner Press, Darien, Connecticut, 1964).
2. Charnov, E. L. *Life History Invariants: Some Explorations of Symmetry in Evolutionary Ecology* (Oxford Univ. Press, Oxford, 1993).
3. Stearns, S. C. *The Evolution of Life Histories* (Oxford Univ. Press, Oxford, 1992).
4. Reiss, M. J. *The Allometry of Growth and Reproduction* (Cambridge Univ. Press, Cambridge, 1989).
5. Ricker, W. E. Growth rates and models. *Fish Physiol.* **8**, 677–743 (1979).
6. von Bertalanffy, L. Quantitative laws in metabolism and growth. *Q. Rev. Biol.* **32**, 217–231 (1957).
7. West, G. B., Brown, J. H. & Enquist, B. J. A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126 (1997).
8. Brown, J. H. & West, G. B. *Scaling in Biology* (Oxford Univ. Press, Oxford, 2000).
9. West, G. B., Brown, J. H. & Enquist, B. J. The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science* **284**, 1677–1679 (1999).
10. Alberts, M. *Molecular Biology of the Cell* (Garland, New York, 1994).
11. Peters, R. H. *The Ecological Implications of Body Size* (Cambridge Univ. Press, Cambridge, 1983).
12. Calder III, W. A. *Size, Function and Life History* (Harvard Univ. Press, Cambridge, Massachusetts, 1984).

13. Rogers, D. M., Olson, B. L. & Wilmore, J. H. Scaling for the $\dot{V}_{O_2}$ -to-body size relationship among children and adults. *J. Appl. Physiol.* **79**(3), 958–967 (1995).
14. Weathers, W. W. & Siegel, R. B. Body size establishes the scaling of avian postnatal metabolic rate: an interspecific analysis using phylogenetically independent contrasts. *Ibis* **137**, 532–542 (1995).
15. Xiaojun, X. & Ruyong, S. The bioenergetics of the southern catfish (*Silurus meridionalis chen*). I. Resting metabolic rate as a function of body weight and temperature. *Physiol. Zool.* **63**, 1181–1195 (1990).
16. Brett, J. R. The relation of size to rate of oxygen consumption and sustained swimming speed of Sockeye Salmon (*Oncorhynchus nerka*). *J. Fish Res. Bd Can.* **22**, 1491–1501 (1989).
17. Hamburger, K. et al. Size, oxygen consumption and growth in the mussel *Mytilus edulis*. *Mar. Biol.* **75**, 303–306 (1983).
18. Enquist, B. J., Brown, J. H. & West, G. B. Scaling of plant energetics and population density. *Nature* **395**, 163–165 (1998).
19. Cummins, K. W. & Wuycheck, J. C. Caloric equivalents for investigations in ecological energetics. *Mitt. Int. Verein. Theor. Angew. Limnol.* **18**, 1–158 (1971).
20. Kohler, A. C. Variations in the growth of Atlantic Cod (*Gadus morhua* L.). *J. Fish. Res. Bd Can.* **21**(1), 57–100 (1964).
21. Blueweiss, L. et al. Relationships between body size and some life history parameters. *Oecologia* **37**, 257–272 (1978).
22. Kozlowski, J. Optimal allocations of resources explains interspecific life-history patterns in animals with indeterminate growth. *Proc. R. Soc. Lond. B* **263**, 559–566 (1996).
23. Day, T. & Taylor, P. D. von Bertalanffy's growth equation should not be used to model age and size at maturity. *Am. Nat.* **149**, 381–393 (1997).
24. Enquist, B. J., West, G. B., Charnov, E. L. & Brown, J. H. Allometric scaling of production and life-history variation in vascular plants. *Nature* **401**, 907–911 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank E. Charnov for discussing the role of reproduction in our formalism, L. Thomson for supplying data on salmon and P. Taylor for comments. Support from the National Science Foundation and the National Center for Ecological Analysis and Synthesis are gratefully acknowledged. J.H.B. and G.B.W. also acknowledge the support of the Thaw Charitable Trust and a Packard Interdisciplinary Science Grant.

Correspondence and requests for materials should be addressed to G.B.W. (e-mail: gwb@lanl.gov).

Hyperpolarization-activated channels HCN1 and HCN4 mediate responses to sour stimuli

David R. Stevens*, Reinhard Seifert†, Bernd Bufe‡, Frank Müller†, Elisabeth Kremmer§, Renate Gauss†, Wolfgang Meyerhof‡, U. Benjamin Kaupp† & Bernd Lindemann*

* Department of Physiology, Universität des Saarlandes, D-66421 Homburg, Germany

† Institute for Biological Information Processing, Forschungszentrum Jülich, D-52425 Jülich, Germany

‡ Department of Molecular Genetics, Deutsches Institut für Ernährungsforschung, Arthur-Scheunert-Allee 114-116, D-14558 Potsdam-Rehbrücke, Germany

§ Institute for Molecular Immunology, GSF, D-81377 München, Germany

Sour taste is initiated by protons acting at receptor proteins or channels. In vertebrates, transduction of this taste quality involves several parallel pathways^{1–5}. Here we examine the effects of sour stimuli on taste cells in slices of vallate papilla from rat. From a subset of cells, we identified a hyperpolarization-activated current that was enhanced by sour stimulation at the taste pore. This current resembled I_h found in neurons and cardio-myocytes^{6,7}, a current carried by members of the family of hyperpolarization-activated and cyclic-nucleotide-gated (HCN) channels^{8–13}. We show by *in situ* hybridization and immunohistochemistry that HCN1 and HCN4 are expressed in a subset of taste cells. By contrast, gustducin, the G-protein involved in bitter and sweet

taste¹⁴, is not expressed in these cells. Lowering extracellular pH causes a dose-dependent flattening of the activation curve of HCN channels and a shift in the voltage of half-maximal activation to more positive voltages. Our results indicate that HCN channels are gated by extracellular protons and may act as receptors for sour taste.

Whole-cell recordings were performed in slices of vallate papilla (Fig. 1a). In 88 out of 497 excitable cells tested (18%), hyperpolarization activated a slowly developing inward current (Fig. 1b), which was blocked by 2 mM extracellular Cs⁺ ($n = 9$) but not intracellular Cs⁺ ($n = 6$). The activation was determined using a dual-pulse protocol (Fig. 1c). The voltage at which activation was half-maximal ($V_{1/2}$) was -105 ± 5 mV ($n = 12$) (Fig. 1d). The time course of activation was voltage dependent and well described by a single time constant, τ , of 478 ± 56 ms at -90 mV ($n = 4$) and 103 ± 14 ms at -130 mV ($n = 6$). The apparent reversal voltage (extrapolated intercept of the instantaneous current–voltage (I – V) relation before (-60 mV) and after (-100 mV) activation) was $E_h = -22.5 \pm 9$ mV (Fig. 1e, f; $n = 4$). These properties are typical of the hyperpolarization-activated current I_h (refs 6, 13).

We challenged 277 of 497 cells by micro-jets of sour stimuli at the taste pore (10 mM citric acid at pH 3.0 or 5.0). In 61 of the 277 cells (22%), the sour stimulus induced an inward current and an increase

in conductance (Fig. 1g). Citrate at pH 6.8 had no effect. Thirty-seven out of sixty-one sour-responsive cells (61%; group 1) also exhibited I_h . Nine cells that showed no I_h before sour stimulation exhibited I_h during sour stimulation (9/61, 15%; group 2). The τ of activation of currents from group 2 was 111 ± 20 ms at -130 mV ($n = 6$). Both group 1 and group 2 cells displayed upregulated I_h during sour stimulation (the response of a group 1 cell is shown in Fig. 1h).

The apparent reversal voltage of the sour-induced current in I_h -positive cells (extrapolated intercept of the instantaneous I – V relations before and during sour stimulation) was $E_{sour} = -15.7 \pm 10$ mV, which is similar to the value of E_h (Fig. 1i). The upregulation of I_h paralleled the increase in membrane conductance and the generation of inward current ($n = 10$; Fig. 1j). Extracellular Cs⁺ (1 mM) blocked the sour-induced inward current by $52 \pm 10\%$ at -70 mV ($n = 7$; Fig. 1k), strengthening the conclusion that I_h contributes to the sour-induced inward current. Of 216 cells not responding to the sour stimulus, only 4 (1.9%) had I_h . In all, 37 out of 41 cells exhibiting I_h (90%, only group 1) or 46 of 50 cells (92%, groups 1 and 2) were sour-responsive. Thus, the correlation is strong between the presence of I_h and the response to the sour stimulus.

Consistent with the presence of I_h in taste cells, messenger RNA

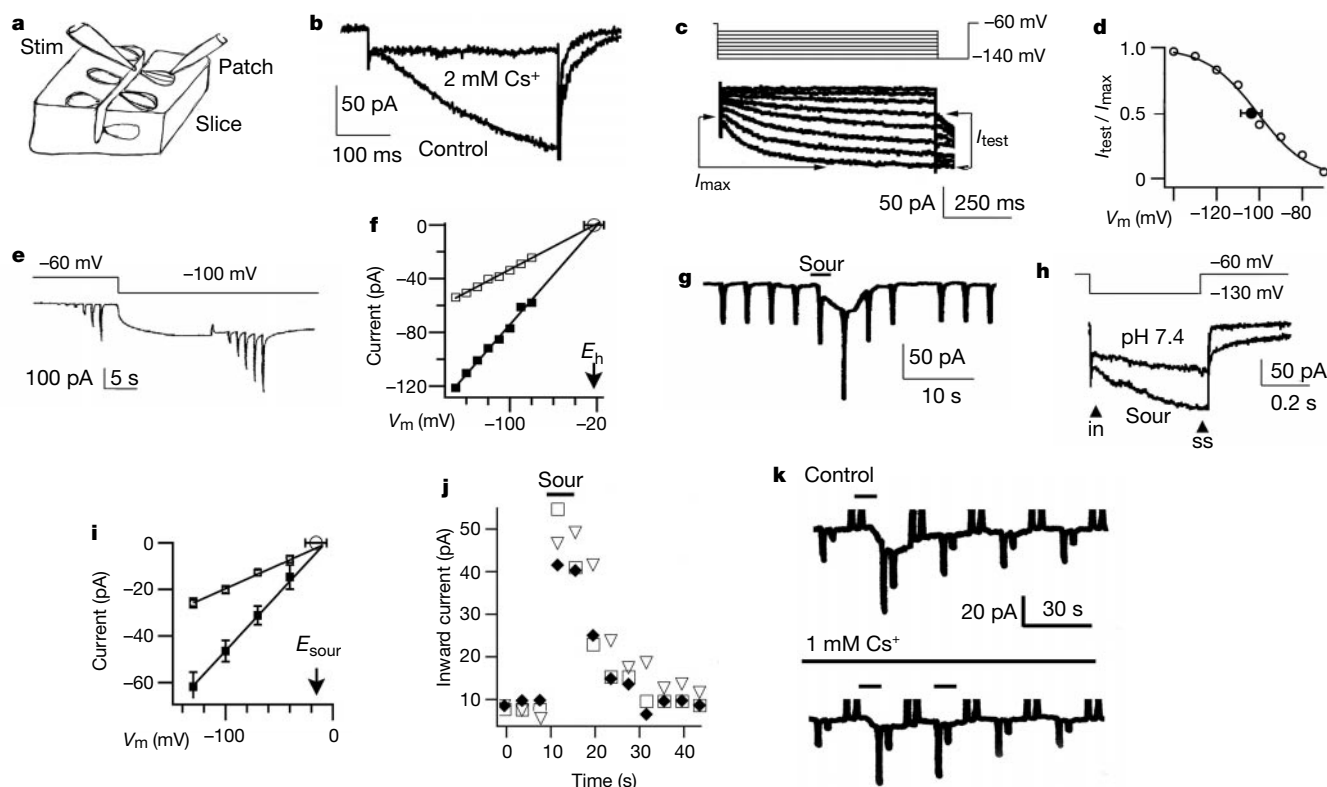


Figure 1 Recordings from rat taste cells. **a**, Slice set-up with micro-jet stimulation (stim) at the taste pore. **b**, Currents activated by stepping from -60 to -130 mV (control), and blockage of this current by 2 mM Cs⁺. **c**, Currents first activated by 870-ms steps from a holding voltage of -60 mV to test voltages from -70 to -140 mV in 10-mV increments. A second step to -140 mV was applied to record tail currents. **d**, Normalized tail currents I_{test}/I_{max} (see **c** for definition) were plotted against test voltage. A fit with the Boltzmann function $1 - I_{test}/I_{max} = (1 + \exp[N_6(V - V_{1/2})/RT])^{-1}$ yielded an apparent $V_{1/2} = -105.5 \pm 5$ mV ($n = 12$) and an apparent $N_6 = 3.8 \pm 0.7$ ($n = 12$). As currents did not reach steady state at all voltages, $V_{1/2}$ is biased towards negative values. **e**, Currents elicited by 300-ms voltage steps from a holding voltage of either -60 or -100 mV to test voltages ranging from -80 to -150 mV in 10-mV increments. **f**, Instantaneous currents as in **e** plotted against test voltage before (open squares) and after (filled squares) activation

by voltage. The intercept indicates a reversal voltage E_h of -22.5 ± 9 mV. **g**, Inward current at -60 mV elicited by application of a sour stimulus (10 mM citrate, pH 5) to the taste pore. Currents elicited by 300-ms steps to -140 mV indicate an increase in conductance of the cell. **h**, Currents activated by a step from -60 to -130 mV at pH 7.4 and pH 5. Instantaneous (in) and steady-state currents (ss) increase during the sour stimulus. **i**, Instantaneous currents from a recording as in **h**, plotted against test voltage before (open squares) and during (filled squares) sour stimulation. The intercept indicates a reversal voltage E_{sour} of -15.7 ± 10 mV ($n = 6$). **j**, Time course of holding current at -60 mV (squares), membrane conductance (steps to -130 mV, triangles) and I_h (diamonds) during a single sour response. **k**, Inward currents at -70 mV elicited by sour stimuli (horizontal bars) in the absence and presence of 1 mM Cs⁺.

for two of the four known isoforms, HCN1 and HCN4, was detected in rat vallate papilla tissue by polymerase chain reaction with reverse transcription (RT-PCR) (Fig. 2a). The coding region of HCN1 obtained by RACE (rapid amplification of overlapping 5' and 3' complementary DNA fragments) was identical to the published sequence¹⁵. *In situ* hybridization using antisense RNA probes showed the presence of HCN mRNA in a subset of taste cells (Fig. 2b). The HCN mRNAs were of low abundance in vallate papilla tissue, because two consecutive PCRs with nested sets of gene-specific primers were required for their identification, and tyramide amplification was necessary to give clear signals in single cells. In sections of 95 taste buds from the vallate papilla, 481 out of 2,565 cells (19%) were positive for HCN, in good agreement with the electrophysiological data.

Rabbit polyclonal and rat monoclonal antibodies specific for HCN1, -2, -3 or -4 were applied to horizontal sections of rat and mouse vallate papilla. In both rat and mouse, antibodies against HCN1 and HCN4 stained a subset of elongated cells in taste buds (Fig. 3a, b), whereas no staining was observed for HCN2 and HCN3. HCN4-like immunofluorescence was concentrated in the plasma membrane (Fig. 3a, arrow). Co-expression of HCN1 and HCN4 was assessed by applying a rabbit antiserum against HCN4 (green) and a rat monoclonal antibody against HCN1 (red) to the same section (Fig. 3b). HCN1 and HCN4 were colocalized (yellow) in most cells. Few cells stained for only one isoform (Fig. 3b, arrows). The localization of HCN-like immunoreactivity at the basolateral but not the apical membrane is compatible with a model in which protons diffuse through tight junctions to reach the pH-sensitive membrane protein^{16,17}.

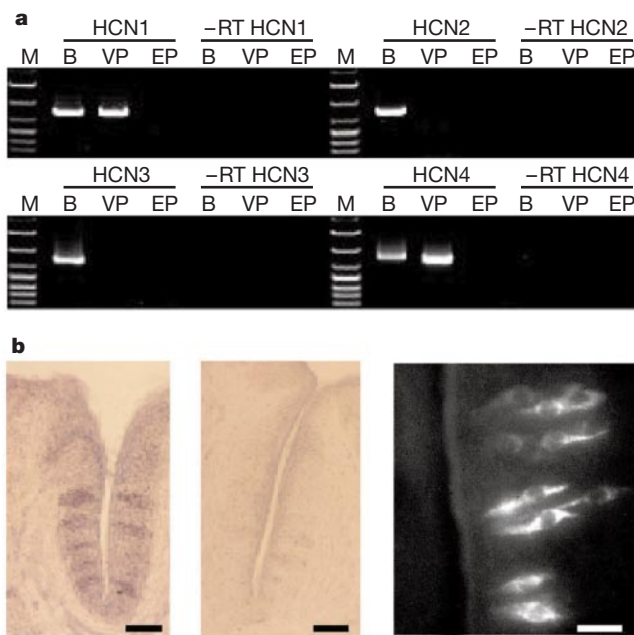


Figure 2 Identification of HCN mRNAs in rat vallate papilla. **a**, Agarose gel electrophoresis of fragments amplified by RT-PCR using subtype-specific primers. mRNAs of HCN1–4 were detected in brain, but only those of HCN1 and HCN4 were found in vallate papilla tissue. The sequence of fragments was identical to the respective sequence of HCN1 and HCN4 from rat brain¹⁵. M, size standard; B, brain (positive control); VP, vallate papilla; EP, non-gustatory tongue epithelium; –RT, no reverse transcriptase present (negative control). Equal amounts of cDNA in the amplification reaction and equal loading of the lanes was verified by RT-PCR of β -actin (not shown). **b**, *In situ* hybridization of rat vallate papilla with HCN RNA probes. Left, hybridized with dioxigenin-labelled antisense RNA; middle, hybridized with sense RNA probes, processed by a standard alkaline phosphatase; right, antisense and tyramide signal amplification. Black bars, 50 μ m; white bar, 20 μ m.

We also labelled sections (Fig. 3c) for both HCN4 (red) and gustducin (green), a G-protein α -subunit that mediates bitter and possibly sweet taste¹⁴. No colocalization was observed, suggesting that sweet- and bitter-responding cells do not express HCN channels. In confocal images taken from eight double-stained sections, we determined the number of HCN-positive cells (70) relative to gustducin-positive cells (94). Given that there are 9.8 gustducin-positive cells per taste bud in the vallate papilla¹⁸, we estimate a value of roughly 7.5 HCN-positive cells per taste bud.

HCN1 and HCN4 currents were studied after heterologous expression of the channels in FLP-IN-293 cells. Figure 4a shows whole-cell HCN1 currents at both pH 7.4 and pH 3.9. The activation threshold at pH 7.4 was near –50 mV and shifted to 0 mV at pH 3.9, showing strong channel activation by protons. Maximal current amplitudes at –120 mV were largely independent of pH. Steady-state inward currents (Fig. 4b) had less rectification at pH 3.9 than at pH 7.4. Figure 4c shows tail currents at –100 mV, from which the voltage dependence of channel activation was determined. In Fig. 4d, mean normalized tail currents are plotted versus membrane voltage V_m for six pH values. Decreasing pH shifted $V_{1/2}$ to less negative voltages and flattened the activation curve. Figure 4e shows the pH dependence of the $V_{1/2}$ and the apparent number of elementary charges (N_e) necessary for activation, as calculated from the Boltzmann equation.

Protons also potentiated HCN4 currents, shifting $V_{1/2}$ and flattening the activation curve (data not shown). However, pH affected the activation kinetics of HCN4 more profoundly than those of HCN1 (Fig. 4f). Decreasing pH from 7.4 to 4.9 led to a strong decrease in the activation time constant τ (Fig. 4g). This acceleration of activation explains the apparent lack of I_h activation in group 2 taste cells at neutral pH. Most likely, these cells express only HCN4 and the voltage pulses used were too short to activate HCN4 appreciably under control conditions (Fig. 4g).

Short pressure applications of sour stimuli to cells expressing HCN1 ($n = 5$; Fig. 4h) and HCN4 ($n = 2$; Fig. 4i) elicited transient

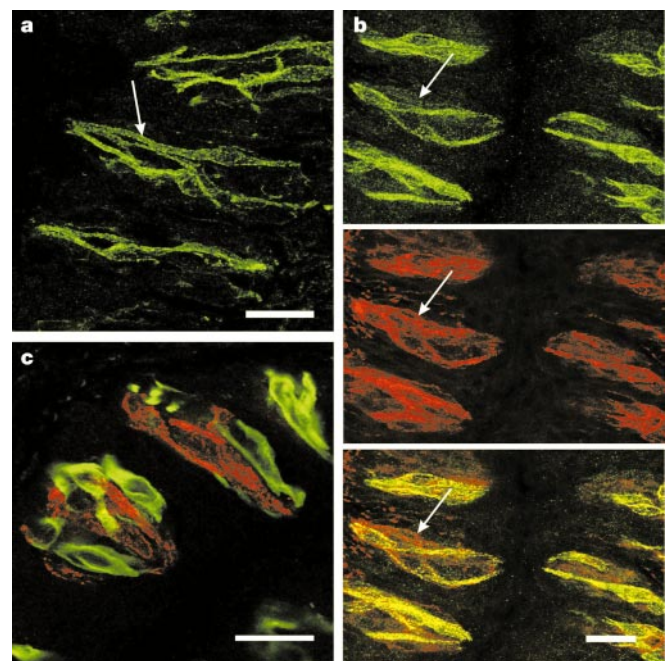


Figure 3 Immunohistochemical localization of HCN1 and HCN4 isoforms. **a**, Horizontal section of mouse vallate papilla, stained for HCN4. Arrow indicates plasma membrane. **b**, In most cases, HCN4 (green) and HCN1 (red) are colocalized; few cells express primarily HCN1 (arrows) or HCN4 (not shown). **c**, Gustducin (green) and HCN4 (red) are expressed in different subsets of cells. Scale bars, 10 μ m.

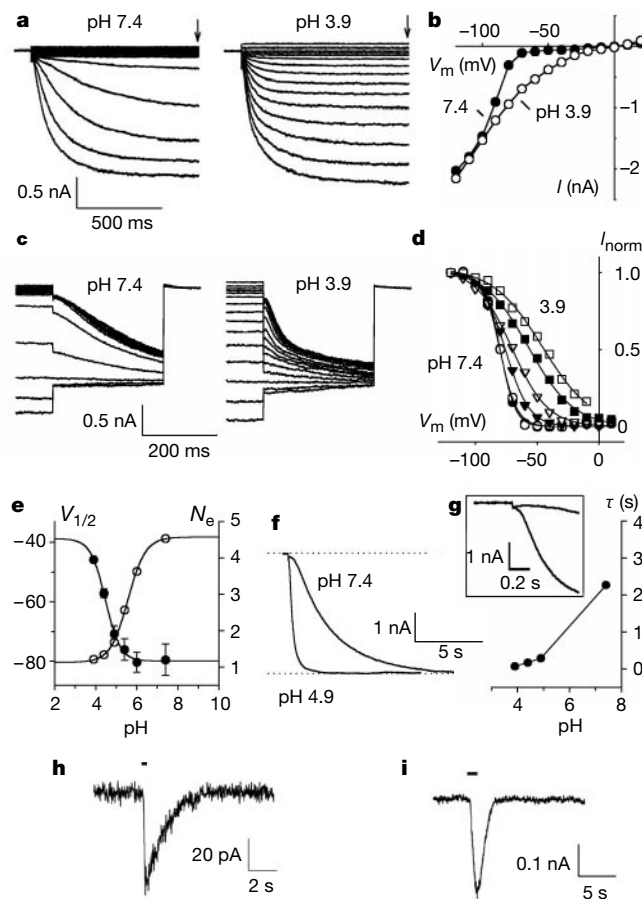


Figure 4 Activation of heterologously expressed HCN1 and HCN4 channels by extracellular protons. **a**, Whole-cell currents of HCN1 channels activated by voltage steps from a holding voltage of 0 mV to various test voltages from +20 to -120 mV in increments of 10 mV at pH 7.4 and at pH 3.9. **b**, Steady-state inward currents of the recordings shown in **a** (arrows), for pH 7.4 (filled circles) and 3.9 (open circles). **c**, Tail currents at -100 mV of the recordings shown in **a**, at pH 7.4 (left) and 3.9 (right). **d**, Voltage dependence of mean normalized tail currents recorded at -100 mV and pH 7.4 (filled circles), 6 (open circles), 5.4 (filled triangles), 4.9 (open triangles), 4.4 (filled squares) and 3.9 (open squares). HCN1 currents were activated for 3 s before tail currents were recorded. Solid lines, Boltzmann functions fitted to data by a least-squares algorithm. Respective values for $V_{1/2}$ and N_e were -79.5 ± 5.3 mV and 4.5 (pH 7.4; $n = 11$); -80.3 ± 3.3 mV and 3.6 (pH 6; $n = 3$); -76.0 ± 3.6 mV and 2.6 (pH 5.4;

$n = 4$); -70.8 ± 2.8 mV and 1.7 (pH 4.9; $n = 6$); -57.1 ± 1.5 mV and 1.3 (pH 4.4; $n = 4$); and -45.8 mV and 1.2 (pH 3.9; $n = 1$). **e**, pH dependence of both $V_{1/2}$ for HCN1 (filled circles) and the number of apparent elementary charges (N_e) necessary to activate HCN1 channels (open circles). **f**, Activation of whole-cell HCN4 currents by a voltage step (14 and 10 s, respectively) from a holding voltage of 0 to -120 mV at pH 7.4 and 4.9. **g**, pH dependence of the activation time constant τ for HCN4. Inset, currents shown in **f** on an expanded timescale. **h**, Rapid response at -50 mV to local application of a sour stimulus in cells expressing HCN1. Bar represents the 0.2-s sour stimulus (pH 4.7) ejected from a patch pipette. **i**, Response at -70 mV to local application to a cell expressing HCN4 channels. Bar represents the 1-s sour stimulus (pH 4.7) ejected from a patch pipette.

inward currents with activation times shorter than 100 ms. The extracellular protons probably interact with a site on the outer surface of the HCN channel, because intracellular acidification causes a negative shift of the activation curve, resulting in inhibition rather than activation of I_h (refs 19, 20).

In conclusion, HCN channels are expressed in a subset of taste receptor cells in the rat vallate papilla. Their currents are activated not only by voltage and cyclic nucleotides, but also strongly and directly by extracellular protons, suggesting that HCN channels function as additional receptors for sour taste. □

Methods

Electrophysiology

The tongue of a juvenile rat (Wistar) was placed for 2 min in ice-cold Tyrode's solution (in mM): 135 NaCl, 5 KCl, 1 CaCl₂, 1 MgCl₂, 10 HEPES, 10 NaHCO₃, 10 D-glucose; pH 7.4. The dissected vallate papilla was sliced in ice-cold Tyrode's solution. The slices (100 μ m) were treated with Tyrode's solution containing collagenase (0.05 mg ml⁻¹; Boehringer), Dispase (0.025 mg ml⁻¹; Boehringer) and trypsin inhibitor (0.05 mg ml⁻¹; Sigma) at room temperature for 7 min. In the experimental chamber, slices were superfused with Tyrode's at a rate of 1–2 ml min⁻¹.

We filled heat-polished borosilicate pipettes (3–6 M Ω) with (in mM) 140 KCl, 0.5 CaCl₂, 2.0 MgCl₂, 5 EGTA, 10 HEPES; pH 7.2. After seal formation, access to the cell interior was achieved by stronger suction combined with a large voltage step. To block I_h , 1–2 mM CsCl was added to the Tyrode's solution for at least 10 min.

The sour stimulus comprised 10 mM citric acid, which replaced 10 mM NaCl in the standard Tyrode's solution. The pH was adjusted to 3.0 with HCl and to either 5.0 or 6.8 with NaOH. The stimulus solution was ejected from a patch pipette placed near the taste pore (Fig. 1a), using a Picospritzer (pulse duration 0.1–20 s; General Valve, USA). We could recognize cases in which the stimulus extended too far, and reached the lower dendrite and the soma of the cell, by a reduction in voltage-dependent Na⁺ or K⁺ currents caused by low pH. Repositioning of the stimulus pipette often resulted in recovery of voltage-dependent currents while sour responses were still elicited. This behaviour suggests that sour stimuli are effective when restricted to the apical cellular pole.

Heterologous expression

The cDNAs for human HCN1 and HCN4 (ref. 21) were subcloned into plasmid pCDN5/FRT (Invitrogen). FLP-IN-293 cells (Invitrogen) were transfected with hHCN1 or hHCN4 to generate stable cell lines. Cells were grown and maintained at 37 °C in DMEM cell-culture medium supplemented with 10% fetal calf serum and 1% antibiotic/antimycotic solution (Gibco). Whole-cell recordings were performed in a high K⁺ medium containing (in mM) 150 KCl, 1.8 CaCl₂, 2.8 MgCl₂, 20 'pH buffer'. The solution at pH 7.4 was buffered with HEPES; all other solutions were buffered with MES. pH was adjusted with HCl or N-methyl-D-glucamine-OH (NMDG). The pipette solution contained (in mM) 150 KCl,

1 MgCl₂, 1 EGTA, 10 HEPES; pH 7.4 (NMDG). Time-resolved activation of HCN currents by protons was studied by ejecting a sour stimulus from a patch pipette with a pressure pulse (Eppendorf microinjector 5242). The extracellular and stimulus solutions for these experiments were (in mM) 120 NaCl, 5 NaOH, 20 KCl, 1 MgCl₂, 1 CaCl₂, 10 HEPES; pH 7.4 or 4.7 (HCl). The intracellular solution contained (in mM) 125 KCl, 25 KOH, 10 EGTA, 10 HEPES; pH 7.4.

RT-PCR and *in situ* hybridization

We prepared RNA from isolated vallate papilla tissue²² and synthesized cDNA using standard techniques (RNA-Clean, AGS; Smart cDNA synthesis Kit, Clontech). Aliquots of vallate papilla cDNA, and rat brain cDNA as a control, were amplified by nested PCR using HCN gene-specific primers. HCN1 (GenBank AF247450): forward, 634–656 and 667–694; reverse, 2,000–2,019 and 1,957–1,979. HCN2 (AF247451): forward, 632–655 and 670–693; reverse, 1,991–2,010 and 1,952–1,973. HCN3 (AF247452): forward, 446–466 and 484–505; reverse, 1,796–1,816 and 1,761–1,780. HCN4 (AF247453): forward, 1,181–1,203 and 1,217–1,239; reverse, 2,537–2,559 and 2,503–2,523. Amplification was carried out for 2 min at 94 °C for 35 cycles, including 1 min at 94 °C, 1 min at the respective annealing temperature and 1.5 min at 68 °C using Advantage 2 polymerase (Clontech). The first reaction (4%) was used as template in the second reaction. Annealing temperatures for the first and second reaction were, respectively, 62 °C and 55 °C (HCN1 and HCN2), 66 °C and 70 °C (HCN3), and 66 °C and 65 °C (HCN4).

Vertical cryo-sections (20 µm) of rat tongues were mounted on microscope slides (Merck Eurolab), dried for 30 min at 50 °C, rehydrated for 2 min in PBS (145 mM NaCl, 1.4 mM KH₂PO₄, 8 mM Na₂HPO₄; pH 7.4), and fixed for 10 min in 4% paraformaldehyde/PBS at 4 °C. After washing and prehybridization, sections were hybridized with 200 ng ml⁻¹ dioxigenin-labelled HCN1 RNA (nucleotides 867–1,578; Boehringer) for 16 h at 65 °C in hybridization solution (50% formamide, 0.75 M NaCl, 75 mM sodium citrate, 0.1% bovine serum albumin, 0.1% polyvinylpyrrolidone, 0.1% Ficoll, 250 µg ml⁻¹ yeast transfer RNA, 500 µg ml⁻¹ herring sperm DNA; pH 7.4). Slides were washed once in 5× SSC for 30 min at 65 °C, and twice in 0.2× SSC for 30 min, 65 °C. To block unspecific reactions during the detection, we incubated sections with blocking solution (1% blocking powder (Boehringer) in 0.1 M maleic acid, 0.15 M NaCl). Detection was either by anti-dioxigenin alkaline phosphatase Fab fragments (Boehringer) (1:2,000 in blocking solution, 4 °C, overnight) or by tyramide signal amplification (NEN) and fluorescein-conjugated avidin D (Linaris).

Immunohistochemistry of HCNs

Vallate papillae of rat (Wistar) and mouse (NMRI) were fixed for 15 min by immersion in 4% paraformaldehyde in 0.1 M phosphate buffer at room temperature. Vallate papillae were washed in phosphate buffer, briefly cryo-protected by incubation in 30% sucrose in phosphate buffer and frozen in OCT compound (Sakura). We stained horizontal cryostat sections (20–30 µm) according to standard immunohistochemistry protocols²³. Omission of the primary antibody abolished the staining. In both rat and mouse, staining was obtained with HCN1- and HCN4-specific antibodies, even though the use of rat monoclonal antibodies in sections of rat tongue yielded high background. Images were taken with a laser scanning confocal microscope (Leica TCS). Antibodies were raised against 35-residue peptides that were specific for each of the four HCN isoforms and were tested for specificity by western blotting and immunocytochemistry on rat and mouse retina (F. Müller *et al.*, manuscript in preparation): IHR-7C3 (rat, 1:50) against HCN1; QQA-4A6 (rat, 1:5) against HCN2; TLL-6C5 (rat, 1:5) against HCN3; PPc73K (rabbit, 1:400), SHG-1E5 (rat, 1:5) against HCN4. The gustducin antibody (rabbit, 1:400) was from Santa Cruz Biotechnology. Anti-rabbit Alexa488 and anti-rat Alexa568 secondary antibodies (both goat, 1:500; Molecular Probes) were used.

Data presentation

Throughout the paper, data are presented as means ± s.d.

Received 31 January; accepted 22 August 2001.

- Gilbertson, T. A., Avenet, P., Kinnamon, S. C. & Roper, S. D. Proton currents through amiloride-sensitive Na channels in hamster taste cells: role in acid transduction. *J. Gen. Physiol.* **100**, 803–824 (1992).
- Kinnamon, S. C., Dionne, V. E. & Beam, K. G. Apical localization of K channels in taste cells provides the basis for sour taste transduction. *Proc. Natl Acad. Sci. USA* **85**, 7023–7027 (1988).
- Ugawa, S. *et al.* Receptor that leaves a sour taste in the mouth. *Nature* **395**, 555–556 (1998).
- Miyamoto, T., Fujiyama, R., Okada, Y. & Sato, T. Sour transduction involves activation of NPPB-sensitive conductance in mouse taste cells. *J. Neurophysiol.* **80**, 1852–1859 (1998).
- Stewart, R. E., Lyall, V., Feldman, G. M., Heck, G. L. & DeSimone, J. A. Acid-induced responses in hamster chorda tympani and intracellular pH tracking by taste receptor cells. *Am. J. Physiol.* **275**, C227–C238 (1998).
- Pape, H. C. Queer current and pacemaker: the hyperpolarization-activated cation current in neurons. *Annu. Rev. Physiol.* **58**, 299–327 (1996).
- Trotier, D. & Doving, K. B. Direct influence of the sodium pump on the membrane potential of vomeronasal chemoreceptor neurones in frog. *J. Physiol. (Lond.)* **490**, 611–621 (1996).
- Santoro, B. *et al.* Identification of a gene encoding a hyperpolarization-activated pacemaker channel of brain. *Cell* **93**, 717–729 (1998).
- Gauss, R., Seifert, R. & Kaupp, U. B. Molecular identification of a hyperpolarization-activated channel in sea urchin sperm. *Nature* **393**, 583–587 (1998).
- Ludwig, A., Zong, X., Jeglitsch, M., Hofmann, F. & Biel, M. A family of hyperpolarization-activated mammalian cation channels. *Nature* **393**, 587–591 (1998).
- Santoro, B. & Tibbs, G. R. The HCN gene family: molecular basis of the hyperpolarization-activated pacemaker channels. *Ann. N. Y. Acad. Sci.* **868**, 741–764 (1999).

- Santoro, B. *et al.* Molecular and functional heterogeneity of hyperpolarization-activated pacemaker channels in the mouse CNS. *J. Neurosci.* **20**, 5264–75 (2000).
- Kaupp, U. B. & Seifert, R. Molecular diversity of pacemaker ion channels. *Annu. Rev. Physiol.* **63**, 235–237 (2001).
- Wong, G. T., Gannon, K. S. & Margolskee, R. F. Transduction of bitter and sweet taste by gustducin. *Nature* **381**, 796–800 (1996).
- Monteggia, L. M., Eisch, A. J., Tang, M. D., Kaczmarek, L. K. & Nestler, E. J. Cloning and localization of the hyperpolarization-activated cyclic nucleotide-gated channel family in rat brain. *Brain Res. Mol. Brain Res.* **81**, 129–139 (2000).
- DeSimone, J. A., Callahan, E. M. & Heck, G. L. Chorda tympani taste response of rat to hydrochloric acid subject to voltage-clamped lingual receptive field. *Am. J. Physiol.* **268**, C1295–C1300 (1995).
- Pittman, D. W. & Contreras, R. J. Responses of single lingual nerve fibers to thermal and chemical stimulation. *Brain Res.* **790**, 224–235 (1998).
- Boughter, J. D. Jr, Pumphrey, D. W., Yu, C., Christy, R. C. & Smith, D. V. Differential expression of α-gustducin in taste bud populations of the rat and hamster. *J. Neurosci.* **17**, 2852–2858 (1997).
- Munsch, T. & Pape, H. C. Modulation of the hyperpolarization-activated cation current of rat thalamic relay neurones by intracellular pH. *J. Physiol. (Lond.)* **519**, 493–504 (1999).
- Zong, X., Stieber, J., Ludwig, A., Hofmann, F. & Biel, M. A single histidine residue determines the pH sensitivity of the pacemaker channel HCN2. *J. Biol. Chem.* **276**, 6313–6319 (2001).
- Seifert, R. *et al.* Molecular characterization of a slowly gating human hyperpolarization-activated channel predominantly expressed in thalamus, heart, and testis. *Proc. Natl Acad. Sci. USA* **96**, 9391–9396 (1999).
- Striemi, B. J., Naim, M. & Lindemann, B. Generation of cyclic AMP in taste buds of the rat circumvallate papilla in response to sucrose. *Cell. Physiol. Biochem.* **1**, 46–54 (1991).
- Müller, F. *et al.* Ligand sensitivity of the α2 subunit from the bovine cone cGMP-gated channel is modulated by protein kinase C but not by calmodulin. *J. Physiol. (Lond.)* **533**, 399–409 (2001).

Acknowledgements

We thank the Deutsche Forschungsgemeinschaft for support.

Correspondence and requests for materials should be addressed to B.L. (e-mail: phblin@uniklinik-saarland.de).

Consequences of a biological invasion reveal the importance of mutualism for plant communities

Caroline E. Christian

Center for Population Biology, One Shields Avenue, University of California, Davis, California 95616, USA

Seed-dispersal mutualisms have a fundamental role in regenerating natural communities^{1,2}. Interest in the importance of seed dispersal to plant communities has been heightened by worldwide declines in animal dispersers^{3–5}. One view, the ‘keystone mutualist hypothesis’, predicts that these human-caused losses will trigger a cascade of linked extinctions throughout the community⁶. Implicitly, this view holds that mutualisms, such as seed dispersal, are crucial ecological interactions that maintain the structure and diversity of natural communities. Although many studies suggest the importance of mutualism^{3,7}, empirical evidence for community-level impacts of mutualists has remained anecdotal^{8,9}, and the central role of mutualism, relative to other species interactions, has long been debated in the theoretical literature^{10,11}. Here I report the community-level consequences of a biological invasion that disrupts important seed-dispersal mutualisms. I show that invasion of South African shrublands by the Argentine ant (*Linepithema humile*) leads to a shift in composition of the plant community, owing to a disproportionate reduction in the densities of large-seeded plants. This study suggests that the preservation of mutualistic interactions may be essential for maintaining natural communities.

Understanding the importance of mutualists to the organization of natural communities is critical for predicting how their decline might alter plant communities. In the fynbos shrublands of South

Africa, where up to 30% of the flora has seeds dispersed by ants, invasion by the Argentine ant provides a test of the community-level importance of mutualism because it results in the elimination of many native ant species that serve as dispersal agents¹².

Ants are vitally important to fynbos plants because they bury seeds in locations that are safe from rodent seed predators and fire¹³. Because seeds are the only stage in the life cycle of most fynbos plants that survives fire, seed burial helps to buffer plant populations from local extinction. Argentine ants do not disperse seeds, and without these services, few unburied seeds survive to germinate after fire¹⁴. Combining controlled field experiments and comparative data from invaded areas, I document a shift in plant community composition in response to an altered assemblage of ant mutualists.

Worldwide, Argentine ants are decimating native-ant faunas in their introduced range^{15–17}. Although many fynbos ant species are eliminated from invaded areas, other ant species coexist with the Argentine ant (Fig. 1 and ref. 18). The consequences of the invasion for fynbos plant communities should depend on the degree to which the remaining native ant species can substitute for the dispersal services lost from invaded areas. The level of substitution will depend on the redundancy of the seed-dispersing partners in the system^{19,20}. In many systems, seed size is a key plant trait that influences disperser assemblages, with larger-seeded species often relying on fewer, more specialized seed dispersers^{21–23}.

To assess whether larger-seeded plants are more vulnerable to the loss of seed dispersers, I tested whether invasion by the Argentine ant differentially affected small- and large-seeded plant species in the Proteaceae. In ant-dispersed members of this family, species produce one of two types of seeds that differ in both seed and reward (elaiosome) size (Fig. 2). Here I focus on a set of co-occurring plant species that represent large- and small-seeded genera. Together, these species comprise 54% (± 10.64 s.d.) of the shrub cover across the study site. Data for each of the large- and small-seeded species were combined into two categories for all analyses (see Methods).

To assess the consequences of the Argentine ant invasion for large- versus small-seeded species, I measured the dispersal efficiency of four common fynbos ant species that are attracted to both types of seeds. Two of these ant species are nearly eliminated by the Argentine ant (*Anoplolepis custodiens* and *Pheidole capensis*) and two

coexist with this invader (*Meranoplus Peringueyi* and *Tetramorium quadrispinosum*). This experiment showed that these native ant species vary in the degree to which they disperse small- and large-seeded Proteaceae.

Although there is some overlap in the seed sizes dispersed by different ant species, two abundant ant species displaced by Argentine ants (*Anoplolepis* and *Pheidole*) are the most effective ants at dispersing large seeds. These ant species dispersed 70% of the large seeds placed outside their nests, compared with only 15% dispersed by ant species that can coexist with the Argentine ant (*Tetramorium* and *Meranoplus*) (Fig. 3). Both groups of ants dispersed small seeds equally well, whereas Argentine ants did not disperse either small or large seeds. This result suggests that plants with larger seeds are more vulnerable to seed predators because the Argentine ant eliminates the most efficient native-ant dispersers of these taxa.

To assess the effects of altered ant communities on seed dispersal and predation, I conducted a factorial experiment in which I either excluded or allowed all ants access to experimental seeds in both invaded and uninvaded areas. This experiment confirmed that elimination of dispersers by the Argentine ant leads to significantly greater vulnerability to seed predation for a large-seeded species, but not for a small-seeded species (Fig. 4). Large seeds placed in invaded areas were less likely to be dispersed by ants and more likely to be consumed by rodents, compared with those in adjacent uninvaded areas. The greater predation risk was not due to underlying differences in seed predation between invaded and uninvaded areas, as seed predation levels were nearly 100% in both areas when ants were excluded from seeds (Fig. 4a). In contrast, seed predation did not differ for small seeds in invaded and uninvaded areas (Fig. 4b). If recruitment is limited by seed survival, these results further support the hypothesis that large-seeded species will be at greater risk of decline in invaded areas than small-seeded species.

To determine whether invasion by the Argentine ant leads to changes in plant community structure, I compared pre-fire and post-fire abundances of plant species in adjacent invaded and uninvaded fynbos vegetation. Because nearly all seed germination occurs in the first season after fire, any differences in post-fire seedling recruitment levels between these areas would reflect the cumulative impacts of the invasion since the last fire cycle at this site (~30 yr ago). Before conducting an experimental burn, I established randomly placed quadrats throughout both invaded and uninvaded areas at the study site and determined the densities of the

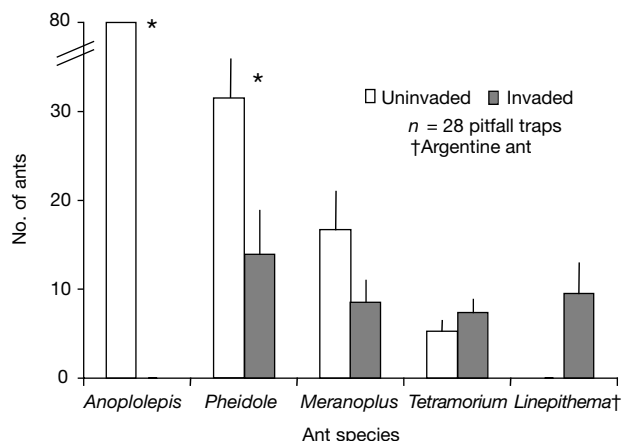


Figure 1 Comparison of the densities of four common, native fynbos ant species and the Argentine ant in invaded and uninvaded areas. Pitfall samples lacking Argentine ants were scored as 'uninvaded' ($n = 28$ pitfall traps). Significant differences in ant densities between invaded and uninvaded areas are noted with an asterisk (univariate analyses of variance (ANOVA) by ant species: *Anoplolepis custodiens* ($F_{1,26} = 3.98$, $P = 0.057$), *Pheidole capensis* ($F_{1,26} = 4.28$, $P = 0.038$), *Meranoplus Peringueyi* ($F_{1,26} = 2.74$, $P = 0.109$), *Tetramorium quadrispinosum* ($F_{1,26} = 3.32$, $P = 0.0799$) and *Linepithema humile* (NA)).

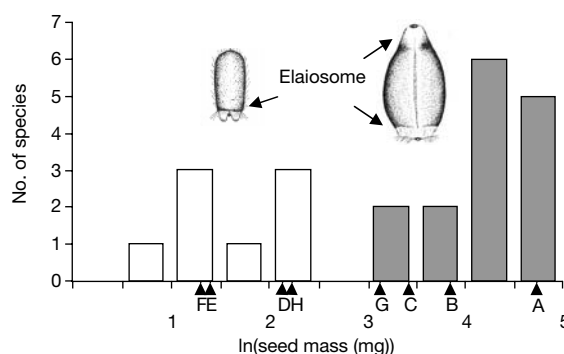


Figure 2 Bimodal distribution of seed size found in ant-dispersed Proteaceae. The mean seed mass (mg) of 23 randomly chosen species of ant-dispersed fynbos Proteaceae is shown on the x axis ($n = 10$ seeds per species). Large-seeded species have rewards (called 'elaiosomes') that encase the seeds, whereas small-seeded species have smaller rewards on one end of the seed. Open bars, small-seeded seed type; filled bars, large-seeded seed type. Arrows indicate the average seed mass of plant species used in the seed-dispersal experiment (A–F) and the comparative study on seedling regeneration (B, C, G, H, D, F). See Methods for species used.

six focal Proteaceae species and all other shrub species in the community.

Pre-burn adult plant densities of the focal species did not differ significantly between the invaded and uninvaded areas (Fig. 5). After the prescribed burn, I re-examined the same quadrat locations for seedlings. To test for differences in patterns of seedling regeneration, I compared changes in seedling recruitment relative to pre-fire adult plant density ($\ln(\text{seedlings (m}^{-2})/\text{adults (m}^{-2}))$). This relative measure of seedling recruitment accounts for any differences in pre-fire adult densities among transects.

Post-fire recruitment of large-seeded Proteaceae was reduced disproportionately in fynbos sites invaded by Argentine ants. As predicted from data on seed dispersal (Fig. 3) and seed predation (Fig. 4a), the average relative seedling density for the three large-seeded species combined was an order of magnitude less in invaded areas than in uninvaded areas (each of the three large-seeded species showed lower recruitment in invaded areas). In contrast, seedling regeneration of two of the three small-seeded species did not differ significantly between invaded and uninvaded sites (Fig. 5).

This significant interaction between the invasion status of a community and the relative densities of different plant species recruiting into the community shows that a change in community composition is occurring within invaded areas. These results probably underestimate the effects of the Argentine ant invasion, as the large-seeded species examined are on the smaller end of the seed-size distribution for this group (Fig. 2, species B, C, G). Because virtually all seedling recruitment occurs within 1 yr after fire in fynbos, post-fire successional changes in plant abundance and dominance arise from differences in plant senescence, rather than from inter-fire seed germination. The low seedling densities of large-seeded Proteaceae in invaded areas cannot be increased by later seed germination and therefore are indicative of lasting community-level changes.

These results and earlier studies^{14,24} indicate that the loss of seed-dispersal services, and its consequent increases in seed predation, is the principal mechanism explaining the reduced densities of large-seeded Proteaceae. Two additional hypotheses could account for the lower densities of these plants observed within invaded areas.

First, Argentine ants might reduce seed production through altered interactions with the insect pollinators. Although Argentine ants have been shown to deter some insect pollinators on *Protea nitida* inflorescences²⁵, they might also deter detrimental insects, such as pre-dispersal seed predators. Although such altered inter-

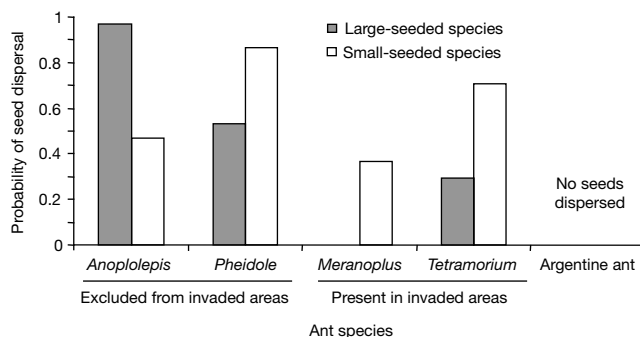


Figure 3 Probability of dispersal for large- and small-seeded Proteaceae by four native fynbos ant species and the Argentine ant. Data from the three focal large-seeded species and the three focal small-seeded species are combined. The interacting effects of seed size and ant species on seed dispersal are demonstrated by a significant three-way interaction (ant species $\times$ seed type $\times$ dispersal fate) in a log-linear analysis ($G^2 = 30.46$, d.f. = 1, $P < 0.0001$). Sixty seeds of each type (large and small) were presented to each ant species.

actions are a possibility, I did not assess their effects on seed production in this study.

Second, seed or seedling survivorship might be lower in invaded areas owing to factors other than predation that lead to reduced seedling densities (such as site differences in abiotic or biotic factors). However, survivorship of experimentally buried seeds and seedling emergence did not differ significantly between adjacent invaded and uninvaded areas for large-seeded *L. truncatulum* (percentage of seed survivorship: $F_{1,90} = 19.57$, $P = 0.165$, $n = 92$ seed cages; percentage of seedling emergence: $F_{1,115} = 0.0054$, $P = 0.9417$, $n = 117$ seed cages; proportion of seedlings surviving: $\chi^2 = 0.072$, degrees of freedom, d.f. = 1, $P = 0.7886$). In summary, the combination of experiments and comparative data presented here indicate that the observed declines in plant densities were due mainly to increased vulnerability to seed predators.

In conclusion, invasion of South African fynbos by the Argentine ant reveals that seed-dispersal mutualisms have a fundamental role in maintaining the structure and diversity of fynbos plant communities. Although the decline of seed dispersers is a growing conservation concern, no studies have assessed the actual responses of plant communities to declining dispersal services. In the case of fynbos Proteaceae, the Argentine ant invasion differentially affects large-seeded species because they rely on fewer, more specialized dispersers for which services are not replaced. These results support

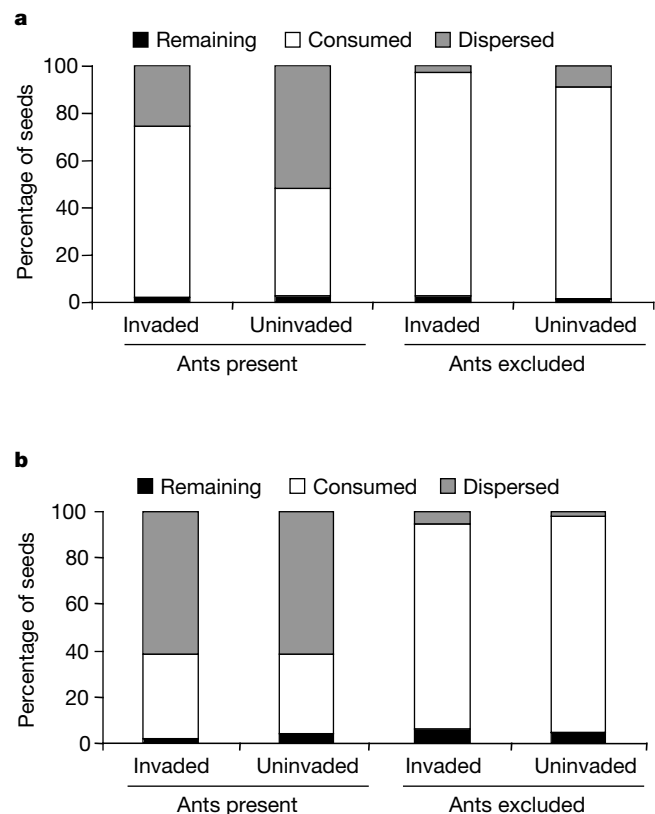


Figure 4 The effects of the Argentine ant on seed predation of large-seeded and small-seeded species. **a**, Large-seeded *Leucospermum truncatulum*; **b**, small-seeded *Spatalla racemosa*. For the large-seeded species (**a**), seed predation levels differed between invaded and uninvaded areas, as illustrated by a significant log-linear test of conditional independence for the invasion status $\times$ seed fate interaction ($G^2 = 73.91$, d.f. = 2, $P < 0.0001$). This test shows that for a given ant-exclusion treatment (yes/no), seed fate was not independent of the invasion status of the community. In contrast, for the small-seeded species (**b**), a log-linear test of conditional independence could not be rejected for the invasion status $\times$ seed fate interaction ($G^2 = 0.54$, d.f. = 2, $P = 0.7634$). For each plant species, 960 seeds were used.

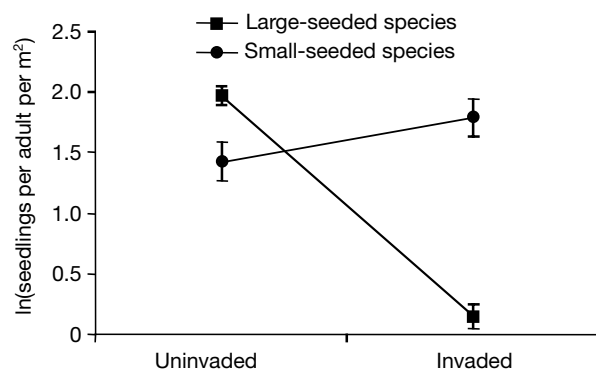


Figure 5 Consequences of the Argentine ant invasion for post-fire seedling recruitment of small-seeded versus large-seeded Proteaceae. Pre-fire adult densities of small-seeded and large-seeded taxa combined did not differ between invaded and uninvaded sites (ANOVA: main effect of invasion status (yes/no) on pre-fire seedling density: $F_{1,234} = 2.45$, $P = 0.3611$). To account for any differences in adult densities, post-fire seedling density was calculated relative to pre-fire densities ($\ln[\text{seedling density (m}^{-2})/\text{adult density (m}^{-2})]$). Data for the three large-seeded and the three small-seeded species were combined for analysis. The invasion $\times$ seed size interaction was highly significant in a nested ANOVA: $F_{1,234} = 71.54$, $P < 0.0001$). Each point in the figure is the mean $\pm$ 1 s.e. for each treatment level of seed-size–invasion status. In total, 240 density measurements were sampled.

the hypothesis that species relying on more-specialized mutualist partners are at a greater risk of decline. A positive relationship between seed-disperser specificity and seed size has been found for other prominent fynbos plant taxa²⁶, and this suggests that many more large-seeded species may also be at risk of decline in this biodiversity hotspot. Although we do not know the importance of seed dispersal in other systems, these results point to the urgency of preserving mutualistic interactions.

These findings highlight the unfolding crisis resulting from the invasion of several ant species throughout the world. Although inconspicuous, ants are ecologically important and abundant in nearly all ecosystems²⁷. When introduced by humans into new landscapes, they can become invasive and have effects that reverberate throughout many levels of biological organization. The extended impacts of the Argentine ant invasion reported here are likely to emerge in other regions where they have been introduced, including Mediterranean-type ecosystems found in Europe and North America. Moreover, several other exotic ant species are spreading around the world (for example, *Solanopsis invicta*, *Pheidole megacephala* and *Wasmannia auropunctata*), damaging both natural and agricultural ecosystems²⁸. Despite the devastating effects of these exotic taxa, ant invasions remain one of the most overlooked aspects of the global biodiversity crisis. □

Methods

Plant species

All plant species used in this study occur together at the study site (Highlands, Kogelberg Biosphere Reserve, 34° 15' S, 19° 05' E). Data shown for the seed-dispersal experiment (Fig. 3) are from the following plant species (mean seed mass (in mg) $\pm$ s.d.; $n = 10$ seeds per species). Large-seeded species: A, *Leucospermum conocarpodendron* (131.57 $\pm$ 15.66); B, *Leucospermum truncatulum* (51.29 $\pm$ 9.01); C, *Mimetes cucullatus* (31.74 $\pm$ 7.18). Small-seeded species: D, *Serruria phyllicoides* (9.02 $\pm$ 2.21); E, *Spatalla racemosa* (4.31 $\pm$ 0.65); F, *Serruria inconspicua* (4.04 $\pm$ 1.27). Recruitment data in Fig. 4 are from species described for Fig. 3, except that (F) *Diastella divaricata* (26.18 $\pm$ 10.39) and (G) *Serruria rubricaulis* (7.85 $\pm$ 1.93) were used in place of *L. conocarpodendron* and *S. phyllicoides* because these latter species did not occur in both invaded and uninvaded sites. Figure 2 shows where each species falls within the seed-size distribution for the ant-dispersed Proteaceae.

Pitfall-trapping study

To determine local densities of native and Argentine ants, I placed 28 pitfall traps along randomly located transects in adjacent invaded and uninvaded fynbos. Pitfall trapping is

an effective method for sampling densities of ground-foraging arthropods, including ants. After 1 month, I collected all pitfall samples from the field, and sorted and counted the ant species.

Seed-dispersal trials

To test whether native ant species vary in their ability to disperse large and small seeds, I placed single seeds of three large-seeded and three small-seeded Proteaceae within 20 cm of 20 nest entrances of each of the four focal ant species and the Argentine ant. All seeds taken into nests within 2 h were recorded as being dispersed (previous studies have shown that seeds remaining above ground for longer than this time are highly vulnerable to seed predators²⁹).

Ant-exclusion experiment

Near each pitfall station, I placed single seeds of a large-seeded species (*Leucospermum truncatulum*) and a small-seeded species (*Spatalla racemosa*) during early morning hours over a 2-week period in peak seed production (December 1999). At half of the seed stations, all ants were excluded from experimental seeds using fluon-coated barriers. In the other half, small windows were cut into fluon barriers to allow ants access to experimental seeds. Rodents were allowed access to all seeds. The ant-exclusion treatment tests for differences in seed predation between invaded and uninvaded areas. The morning after placing out seeds, the fate of each seed was assigned to one of the following categories: (1) remaining at site; (2) consumed by seed predators; or (3) dispersed by ants. In this system, seed-eating rodents (such as *Acomys subspinosus* and *Rhabdomys pumilio*) discard seed coats on site after eating seed embryos³⁰. The presence of empty seed coats allowed me to distinguish seed predation by vertebrates from seed-dispersal events by ants.

Seed-cage experiment

To test for differences in seed and seedling survivorship in adjacent invaded and uninvaded areas, I planted experimental seeds of large-seeded *Leucospermum truncatulum* throughout the study site. All seeds were planted at a uniform depth (3 cm) in wire mesh cages ($n = 9$ seeds per cage; $n = 120$ cages). After a prescribed burn (April 1999), I monitored seedling emergence and seedling survivorship until the end of the summer (March 2000). To estimate seed survivorship of ungerminated seeds, I excavated a subset of the seed cages ($n = 92$ cages). Data on the percentage of seeds surviving and emerging from each seed cage were arcsin transformed for analyses.

Received 28 June; accepted 28 August 2001.

- Howe, H. F. & Smallwood, J. Ecology of seed dispersal. *Ann. Rev. Ecol. Syst.* **13**, 201–228 (1982).
- Chapman, C. A. Primate seed dispersal: coevolution and conservation implications. *Evol. Anthropol.* **4**, 74–82 (1995).
- Cox, P. A., Elmqvist, T., Pierson, E. D. & Rainey, W. E. Flying foxes as strong interactors in South Pacific island ecosystems: a conservation hypothesis. *Conserv. Biol.* **5**, 448–454 (1991).
- Chapman, C. A. & Chapman, L. J. Survival without dispersers: seedling recruitment under parents. *Conserv. Biol.* **9**, 675–678 (1995).
- Aparecida Lopez, M. & Ferrari, S. F. Effects of human colonization on the abundance and diversity of mammals in eastern Brazilian Amazonia. *Conserv. Biol.* **14**, 1658–1665 (2000).
- Gilbert, L. E. In *Conservation Biology: An Evolutionary–Ecological Perspective* (eds Soule, M. E. & Wilcox, B. A.) 11–33 (Sinauer, Sunderland, 1980).
- Bond, W. J. Do mutualisms matter? Assessing the impact of pollinator and disperser disruption on plant extinction. *Phil. Trans. R. Soc. Lond. B* **344**, 83–90 (1994).
- Mills, L. S., Soule, M. E. & Doak, D. F. The keystone-species concept in ecology and conservation. *Bioscience* **43**, 219–224 (1993).
- Power, M. E. et al. Challenges in the quest for keystones. *Bioscience* **46**, 609–620 (1996).
- May, R. M. In *Theoretical Ecology, Principles and Applications* (ed. May, R. M.) 49–70 (Saunders, Philadelphia, 1976).
- Adicott, J. F. In *A New Ecology: Novel Approaches to Interactive Systems* (eds Price, P. W., Slobodkinoff, C. N. & Gaud, W. S.) 437–455 (Wiley, New York, 1984).
- de Kock, A. E. & Giliomee, J. H. A survey of the Argentine ant, *Iridomyrmex humilis* (Mayr) (Hymenoptera: Formicidae) in South African fynbos. *J. Entomol. Soc. S. Africa* **52**, 151–164 (1989).
- Bond, W. J., Yeaton, R. & Stock, W. D. In *Ant–Plant Interactions* (eds Huxley, C. & Cutler, D.) 448–462 (Oxford Univ. Press, Oxford, 1991).
- Bond, W. J. & Slingsby, P. Collapse of an ant–plant mutualism: the Argentine ant (*Iridomyrmex humilis*) and myrmecochorous Proteaceae. *Ecology* **65**, 1031–1037 (1984).
- Cole, F. R., Medeiros, A. C., Loope, L. L. & Zuehlke, W. W. Effects of the Argentine ant on arthropod fauna of Hawaiian high-elevation shrubland. *Ecology* **73**, 1313–1322 (1992).
- Ward, P. S. Distribution of the introduced Argentine ant (*Iridomyrmex humilis*) in natural habitats of the lower Sacramento Valley and its effects on the indigenous ant fauna. *Hilgardia* **55**, 1–16 (1987).
- Way, M. J., Cammell, M. E., Paiva, M. R. & Collingwood, C. A. Distribution and dynamics of the Argentine ant *Linepithema* (*Iridomyrmex*) *humile* (Mayr) in relation to vegetation, soil conditions, topography and native ant competitor ants in Portugal. *Insectes Sociaux* **44**, 415–433 (1997).
- De Kock, A. E., Giliomee, J. H., Pringle, K. L. & Majer, J. D. In *Fire in South African Mountain Fynbos* (eds van Wilgen, B. W., Richardson, D. M., Kruger, F. J. & van Hensbergen, H. J.) 201–215 (Springer, Berlin, 1992).
- Lawton, J. H. & Brown, V. K. In *Biodiversity and Ecosystem Function* (eds Schulze, E.-D. & Mooney, H. A.) 255–270 (Springer, Berlin, 1993).
- Cushman, J. H. In *Islands: Biological Diversity and Ecosystem Function* (eds Vitousek, P. M., Loope, L. L. & Adersen, H.) 135–147 (Springer, Berlin, 1995).
- Brown, J. H. & Henske, E. J. Control of a desert–grassland transition by a keystone rodent guild. *Science* **250**, 1705–1707 (1990).
- Howe, H. F. Specialized and generalized dispersal systems: where does the ‘the paradigm’ stand? *Vegetatio* **107/108**, 3–13 (1993).
- Hughes, L. et al. Predicting dispersal spectra: a minimal set of hypotheses based on plant attributes. *J. Ecol.* **82**, 933–950 (1994).

24. Slingsby, P. & Bond, W. Ants—friends of the Fynbos: *Mimetes hirtus*, dispersal and storage by ants. *Veld Flora* **67**, 39–45 (1981).
25. Visser, D., Wright, M. G. & Gilmore, J. H. The effect of the Argentine ant, *Linepithema humile* (Mayr) (Hymenoptera: Formicidae), on flower-visiting insects of *Protea nitida* Mill. (Proteaceae). *Afr. Entomol.* **4**, 285–287 (1996).
26. Witt, A. B. R. *Factors affecting Myrmecochory in Fynbos*. (Thesis, Univ. Stellenbosch, 1993).
27. Hölldobler, B. & Wilson, E. O. *The Ants* (Harvard Univ. Press, Cambridge, 1990).
28. Passera, L. in *Exotic ants: Biology, Impact, and Control of Introduced Species* (ed. Williams, D. H.) 23–43 (Westview, Boulder, 1994).
29. Slingsby, P. & Bond, W. J. The influence of ants on the dispersal distance and seedling recruitment of *Leucospermum conocarpodendron* (L.) Buek (Proteaceae). *S. Afr. J. Bot.* **51**, 30–34 (1985).
30. Bond, W. J. & Breytenbach, G. J. Ants, rodents and seed predation in Proteaceae. *S. Afr. J. Zool.* **20**, 150–154 (1985).

Acknowledgements

Comments from A. Agrawal, E. Baack, E. Bruna, H. Cushman, A. Pauw, J. Rudgers, M. Stanton, S. Strauss and H. Robertson improved this manuscript. I thank E. Baack, W. Bond, D. Doak, M. and A. Johns, and P. Ward for their assistance and discussions during this project. Cape Nature Conservation and the South African Museum also provided support. This work was supported by grants from the National Science Foundation, Achievement Rewards for College Scientists Foundation, and the Center for Population Biology, UC Davis.

Correspondence and requests for materials should be addressed to the author (e-mail: cechristian@ucdavis.edu).

***In situ* class switching and differentiation to IgA-producing cells in the gut lamina propria**

Sidonia Fagarasan, Kazuo Kinoshita, Masamichi Muramatsu, Koichi Ikuta & Tasuku Honjo

Department of Medical Chemistry, Graduate School of Medicine, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto, 606-8501, Japan

One of the front lines of the immune defence is the gut mucosa, where immunoglobulin- α (IgA) is continuously produced to react with commensal bacteria and dietary antigens. It is generally accepted that, after antigenic stimulation in the Peyer's patches, IgA⁺ lymphoblasts (B220⁺IgA⁺) migrate through the lymph and blood circulation, and eventually home to the lamina propria of the intestine^{1,2}. Mice that lack activation-induced cytidine deaminase (AID) are defective in class switch recombination (CSR) and somatic hypermutation³. CSR changes the immunoglobulin heavy chain constant region (C_H) gene being expressed from C μ to other C_H genes, resulting in a switch of the immunoglobulin isotype from IgM to IgG, IgE or IgA. AID^{-/-} mice also secrete large amounts of immunoglobulin- μ (IgM) into faeces, and accumulate B220⁺IgM⁺ plasma cells as well as B220⁺IgM⁺ cells in the gut. Here we show that lamina propria B220⁺IgA⁺ cells have just completed CSR, as they still express both AID and transcripts from circular DNA that has been 'looped-out' during CSR. Lamina propria IgM⁺ B cells seem to be pre-committed to switching to IgA⁺ *in vitro* as well as *in vivo*. Culturing lamina propria IgM⁺ B cells together with lamina propria stromal cells enhances preferential switching and differentiation of B cells to IgA⁺ plasma cells. We conclude that IgA⁺ cells in the gut lamina propria are generated *in situ* from B220⁺IgM⁺ lymphocytes.

We found that AID^{-/-} mice secrete increasing amounts of IgM in faeces with age (Fig. 1a). This is probably because in the lamina propria (LP) of AID^{-/-} mice a population of B220⁺IgM⁺ cells (Fig. 1b) with intracytoplasmic IgM and plasma cell morphology (Fig. 1c) increases with age (Fig. 1d). We also found that B220⁺IgM⁺

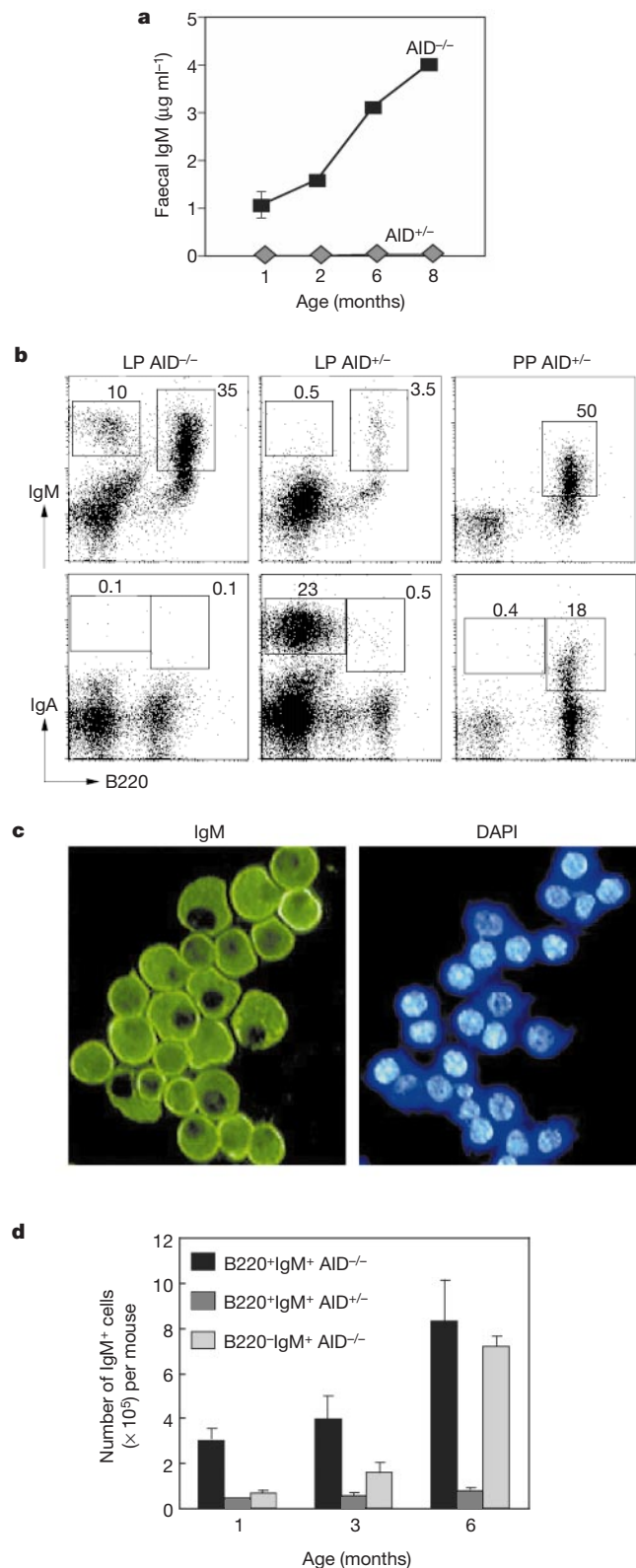


Figure 1 Accumulation of B220⁺IgM⁺ plasma cells and B220⁺IgM⁺ B cells in the LP of aged AID^{-/-} mice. **a**, IgM levels in faeces of AID^{-/-} mice determined by ELISA. **b**, Flow cytometric analysis of LP cells from AID^{-/-} mice, and LP and PP cells from AID^{+/-} mice, stained for B220 and IgM or IgA. Percentages of gated populations are shown. **c**, Sorted B220⁺IgM⁺ cells from the LP of AID^{-/-} mice, stained for intracytoplasmic IgM and counterstained with 4'6'-diamidino-2-phenylindole (DAPI) to visualize nuclei. **d**, Total numbers of B220⁺IgM⁺ and B220⁺IgM⁺ plasma cells in the LP of AID^{-/-} mice, and B220⁺IgM⁺ cells in the LP of AID^{+/-} at the indicated months, calculated by (percentage of cells) × (number of viable cells).

lymphocytes accumulate in the LP of AID^{-/-} mice with age (Fig. 1b, d). By contrast, the LP of normal and AID^{+/-} mice contained a large number of B220⁺IgA⁺ plasma cells⁴, and a small but significant number of B220⁺IgA⁺ B cells ($2.2 \pm 0.16 \times 10^4$ per mouse; Fig. 1b). Note that B220⁺IgM⁺ cells are more abundant than B220⁺IgA⁺ cells in the LP (Fig. 1b). Peyer's patches (PPs) contained several B220⁺IgA⁺ cells ($19.5 \pm 6 \times 10^4$ per mouse; Fig. 1b), most of which were dividing cells located mainly within the germinal centres, and thus considered to be generated *in situ*⁵. Together, these results suggest that in the LP of normal mice B220⁺IgA⁺ cells

may be generated *in situ* from B220⁺IgM⁺ cells, but that IgM⁺ cells accumulate in the LP of AID^{-/-} mice, probably because these mice lack CSR.

To test this hypothesis, we used three molecular markers for CSR from the μ -chain to the α -chain: (1) AID, which is essential for CSR and strictly induced in stimulated B cells undergoing class switching^{3,6}; (2) α -germline transcripts (α GTs), production of which is prerequisite for CSR⁷; (3) I α -C μ circle transcripts (α CTs) produced from circular DNA that is looped out and lost after CSR⁸ (Fig. 2a). α CT was expressed in switch-induced but not

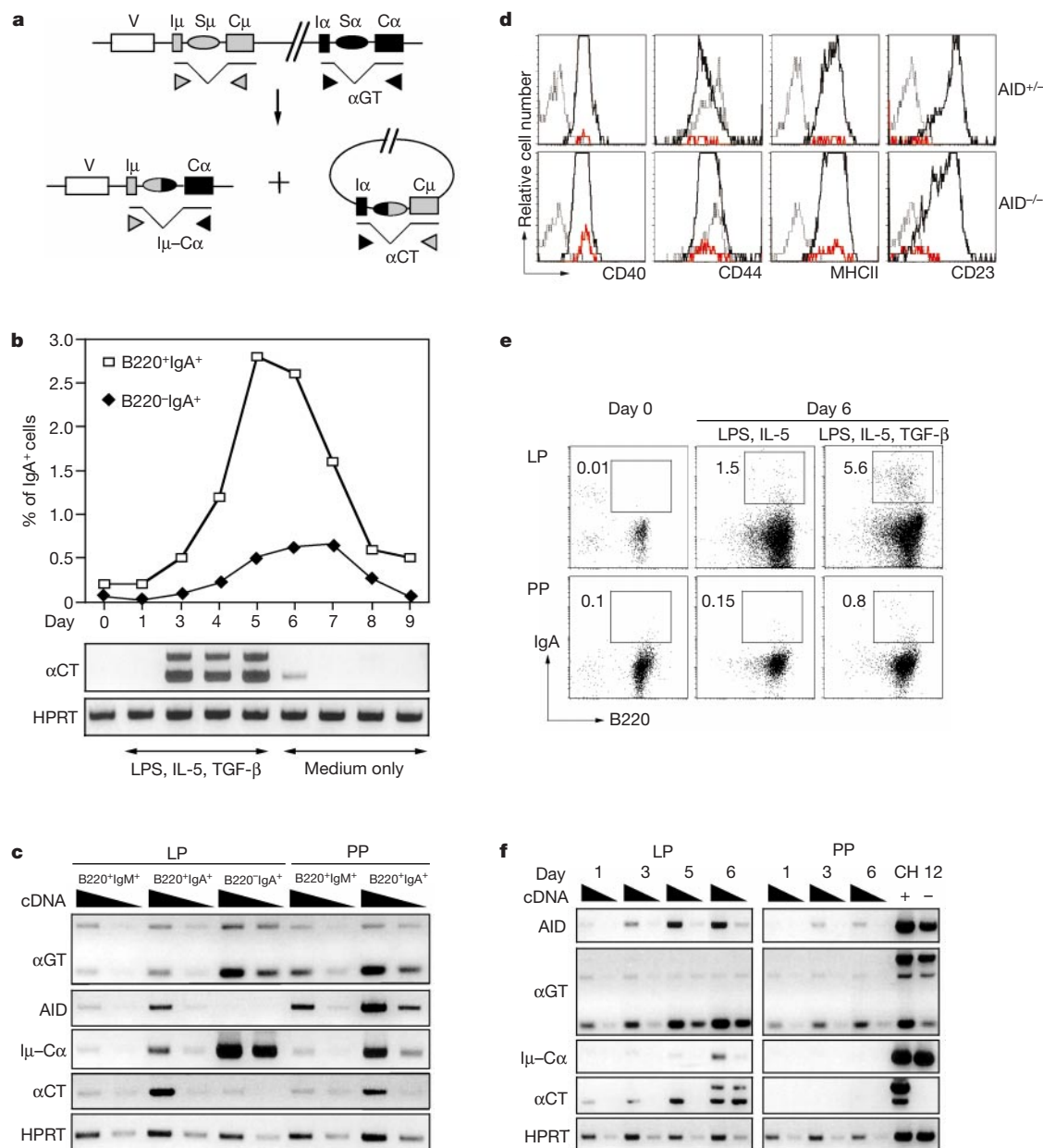


Figure 2 Active CSR in B220⁺IgA⁺ cells of the LP. **a**, Generation of α GT, I μ -C α and α CT transcripts on CSR. **b**, Time-course study of α CT expression (RT-PCR) and surface IgA (FACS) in IgA-depleted splenic B cells stimulated as indicated. **c**, RT-PCR of α GT, AID, I μ -C α and α CT transcripts in LP B220⁺IgM⁺, B220⁺IgA⁺ and B220⁻IgA⁺ cells, and in PP B220⁺IgM⁺ and B220⁺IgA⁺ cells. **d**, LP and PP cells from AID^{+/-} (top) and AID^{-/-} mice (bottom) were stained for B220 and IgM, in combination with the indicated antigens. Solid red and black lines represent LP and PP cells, respectively, both gated on B220⁺IgM⁺ cells. Thin dotted lines represent PP B220⁺IgM⁻ cells. **e**, LP and PP B220⁺IgM⁺ cells were prepared and cultured as described in Methods. Cells were stained for B220 and IgA, and

dead cells were excluded by propidium-iodide gating. Percentages of IgA⁺ cells are shown. Representative data from three different experiments are shown. **f**, RT-PCR of transcripts indicates IgA class switch after *in vitro* culture. Sorted B220⁺IgM⁺ cells from the LP and PPs were cultured with LPS, IL-5 and TGF- β . On the indicated days, cells were collected and RNA was prepared. Transcripts produced in CH12F3 cells, before (-) and 24 h after (+), switch stimulation are shown as controls. Three bands of α GT and two bands of α CT represent alternatively spliced forms. In **c** and **f**, hypoxanthine phosphoribosyltransferase (HPRT) was used as a cDNA control. Triangles indicate a fivefold dilution of cDNA. For experiments shown in **b**, **c**, **e**, **f**, cells were from BALB/c mice.

in non-induced CH12F3 cells⁹ (Fig. 2f). In a time-course study of IgA-depleted spleen cells undergoing class switching after *in vitro* activation, α CT was detected shortly before the appearance of B220⁺IgA⁺ cells and quickly disappeared after switch stimulation was removed (Fig. 2b). Therefore, the presence of α CT is the most decisive hallmark for active CSR, as described elsewhere in more detail.

We isolated IgM⁺ and IgA⁺ B cell populations from the LP and PPs of normal mice, and examined the expression of the molecular markers described above (Fig. 2c). LP B220⁺IgA⁺ cells contained α GT, AID and α CT transcripts, indicating that at least some of these cells had undergone CSR very recently. A similar expression pattern was observed in PP B220⁺IgA⁺ cells. LP B220⁺IgA⁺ plasma cells expressed α GT, but none or very low amounts of AID and α CT, which were quickly downregulated after CSR. All IgA⁺ populations contained I μ -C α transcripts that were expressed from the IgH locus after μ -chain to α -chain CSR (Fig. 2a, c). Together, these observations indicate that CSR is likely to have occurred recently in B220⁺IgA⁺ cells of both the LP and PPs.

LP and PP B220⁺IgM⁺ cells expressed α GT and AID transcripts, albeit at lower levels than their B220⁺IgA⁺ counterparts. This suggests that at least some B220⁺IgM⁺ cells were activated to make their S α region accessible to CSR recombinase and are thus ready to switch to IgA. LP and PP B220⁺IgM⁺ lymphocytes had a similar surface phenotype, expressing MHC class II molecules, CD40 and CD44; however, there was no expression of CD23 (IgE FcR) on LP cells (Fig. 2d). CD23 is expressed on all mature IgM⁺IgD⁺ B cells, but not on marginal-zone B cells and peritoneal B1 cells, and disappears after switch activation and differentiation^{10–13}. Thus, LP B220⁺IgM⁺ cells seem to have characteristics that distinguish them from their PP counterparts, probably owing to the different LP microenvironment.

To test the switch capacity of LP lymphocytes, we isolated

B220⁺IgM⁺ cells and cultivated them under various conditions. Stimulation with lipopolysaccharide (LPS), interleukin (IL)-5 and transforming growth factor- β (TGF- β) for 6 days was most effective in inducing class switching of LP B220⁺IgM⁺ cells to B220⁺IgA⁺ (Fig. 2e, top). We found a marked upregulation of AID and α GT expression in LP B220⁺IgM⁺ cells after day 3, followed by an increase of α CT at day 5 and expression of I μ -C α transcripts at day 6 (Fig. 2f). By contrast, PP B220⁺IgM⁺ cells could not generate IgA-producing cells under similar conditions (Fig. 2e, bottom). The lower levels of AID and α GT induced in PP B220⁺IgM⁺ cells did not seem to be sufficient for an efficient class switching, as shown by the low expression of IgA, as well as of α CT and I μ -C α transcripts (Fig. 2b, f). When PP B cells were cultured in the presence of T cells and other accessory cells, however, their IgA switch efficiency was similar or higher than that of IgA-depleted spleen cells (data not shown).

To confirm the switching capacity of LP B cells *in vivo*, we injected IgA-depleted LP cells from normal mice into RAG-2^{-/-} mice. Similar experiments were done with PP or mesenteric lymph node (MLN) cells. Although roughly the same number of B220⁺IgM⁺ B cells was injected in every experiment, mice injected with LP B cells showed a faster increase in faecal levels of IgA than mice injected with B cells from PPs or MLNs (Fig. 3a), suggesting that LP B cells may contain more IgA-committed cells. At 6 weeks after transfer of LP cells, the most abundant, switched B cells recovered were of IgA isotype (Fig. 3b), and were located mainly in the LP and MLNs of RAG-2^{-/-} mice (Fig. 3c, top). Although at this time point PP B cells reconstituted IgA⁺ populations (Fig. 3c, top), they generated IgG⁺ B cells even more efficiently (Fig. 3b, c, middle).

We confirmed that LP B220⁺IgM⁺ cells from AID^{-/-} mice have a strong IgA class switching capacity by overexpressing AID in these cells by retroviral transformation. Under this condition, LP B cells of

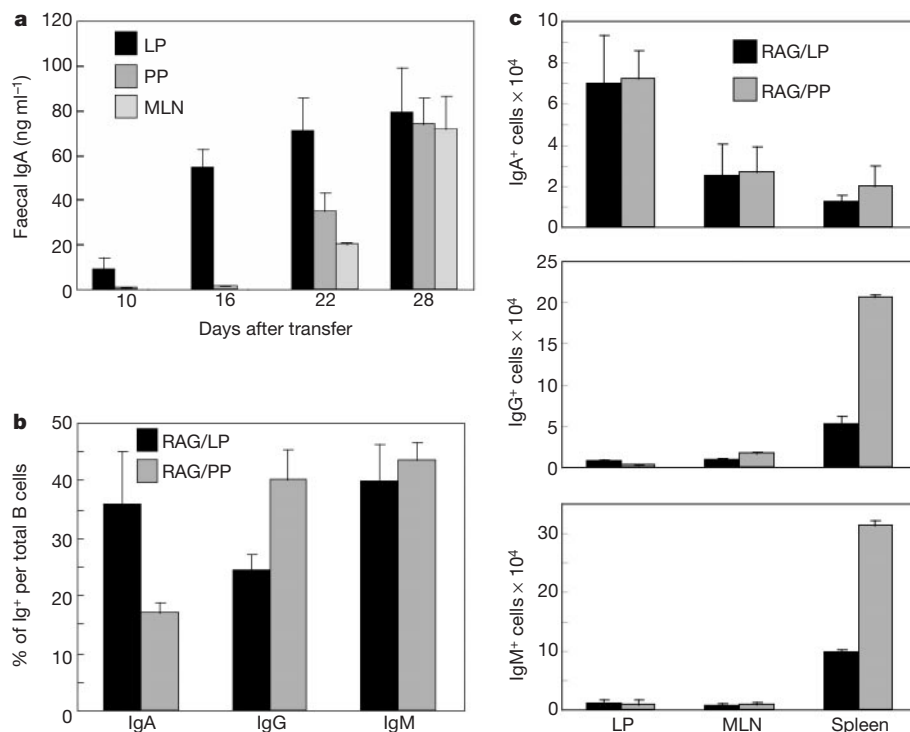


Figure 3 *In vivo* LP IgM⁺ B cells generate IgA-producing cells rapidly and efficiently. **a**, IgA levels in faecal extracts of RAG-2^{-/-} mice at the indicated time points after transfer of IgA-depleted cells from the LP, PPs and MLNs. Results represent mean $\pm$ s.e. from three experiments. **b**, Percentages of IgA⁺, IgG⁺ or IgM⁺ cells per total number of B cells

recovered from the LP, MLNs and spleen of RAG-2^{-/-} mice 6 weeks after transfer of LP and PP cells. **c**, Tissue distribution and cell numbers of IgA⁺, IgG⁺ or IgM⁺ B cells in mice injected with LP and PP cells, calculated by (percentage of Ig⁺ cells) $\times$ (number of viable cells), 6 weeks after cell transfer.

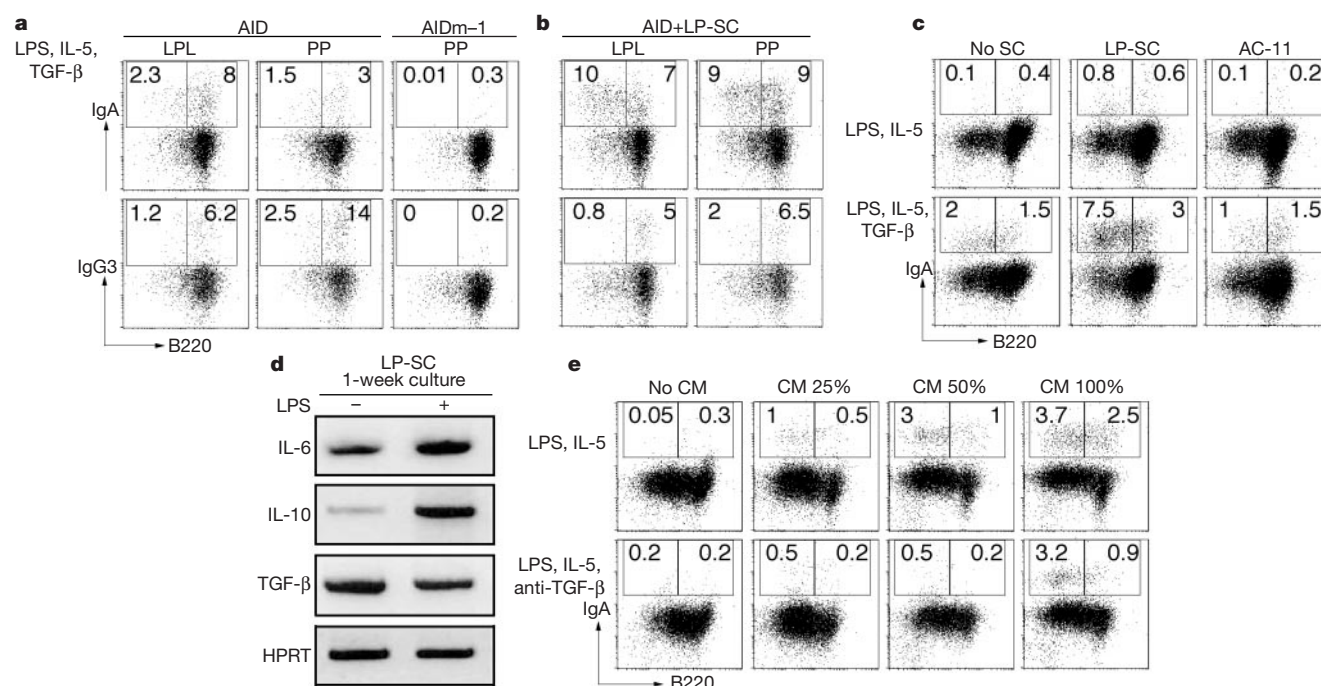


Figure 4 LP-SCs support preferential switching and differentiation to IgA⁺ plasma cells. **a**, LP and PP B220⁺IgM⁺ B cells from AID^{-/-} mice after retroviral transformation were cultured and stimulated as indicated for 5 days. Right, PP B cells overexpressed with a loss-of-function mutant AID (AIDm-1). **b**, LP and PP B220⁺IgM⁺ B cells from AID^{-/-} mice after retroviral transformation were stimulated for 2 days as indicated in **a**, and further cultured with LP-SCs for 3 days under the same stimulation conditions. For experiments shown in **a** and **b**, cells were stained for B220 and IgA (top) or IgG3 (bottom). Numbers represent percentages in gated regions. All cells are positive for green fluorescent protein.

c, Spleen B cells from AID^{-/-} mice were sorted and cultured for 5 days as indicated, on no SCs (left), on LP-SCs (middle) and on bone marrow SCs (right). **d**, RT-PCR of LP-SCs after a 1-week culture without or with LPS (20 μ g ml⁻¹). **e**, Spleen B cells from AID^{-/-} mice were sorted and cultured for 5 days as indicated, without or with conditioned media (CM) from a 4-day culture supernatant, in the presence of the indicated stimulants and antibodies. Percentages of B220⁺IgA⁺ B cells and B220⁺IgA⁺ plasma cells are shown for **c** and **e**.

AID^{-/-} mice generated more IgA⁺ cells than PP B cells when appropriately stimulated (Fig. 4a, top left), although the latter contained many activated germinal centre B cells³. In contrast, PP B cells switched to IgG3 with higher efficiency than LP B cells under the same conditions (Fig. 4a, bottom left) or when stimulated with LPS and IL-5 only (data not shown). As a control, overexpressing a loss-of-function mutant (AIDm-1) of AID did not induce any CSR (Fig. 4a, right). These results indicate that at least some LP IgM⁺ B cells may be committed to switch to IgA.

As there are no clear germinal centres in the LP, there must be another type of microenvironment that can facilitate switching to IgA. Because LP lymphocytes are in close contact with stromal cells (LP-SCs), we tested the capacity of LP-SCs to support class switching and differentiation to plasma cells *in vitro*. We found that SCs isolated from the LP enhanced preferential IgA class switching equally in LP, PP, MLN or spleen B cells from AID^{-/-} mice after overexpression of AID and stimulation with LPS, IL-5 and TGF- β (Fig. 4b, top; and data not shown). The increased switching to IgA was paralleled by a decreased switching to IgG3 (Fig. 4b, bottom).

LP-SCs seemed to enhance not only IgA switching but also differentiation to IgA⁺ plasma cells, as shown by an increase in the B220⁺IgA⁺ population (Fig. 4b, top). In the presence of LPS, IL-5 and TGF- β , LP-SCs but not bone-marrow-derived SCs (AC-11)¹⁴ supported preferential switching and differentiation to IgA⁺ plasma cells of spleen B cells from AID^{-/-} mice (Fig. 4c). LP-SC-induced switching was independent of the interaction of CD40 and CD40L, as LP-SCs do not express CD40L (data not shown). In fact, IgA class switching has been shown to occur *in vivo* in the absence of CD40L in CD40L^{-/-} mice^{15,16}.

LP-SCs contained transcripts for TGF- β , IL-10 and IL-6, which are thought to be important for IgA switching or differentiation to

plasma cells (Fig. 4d)¹⁷⁻¹⁹. Indeed, soluble factors secreted by LP-SCs could induce switching to IgA (Fig. 4e, top); this was achieved mostly through TGF- β , as anti-TGF- β antibodies abolished or diminished the switching effect (Fig. 4e, bottom). LP-SC soluble factors also supported plasma cell differentiation (Fig. 4e), which was probably not caused by IL-6 as an anti-IL-6 antibody did not affect the differentiation capacity (data not shown). IL-10 was a negative rather than a positive regulator of IgA class switching, as addition of IL-10 in culture suppressed the IgA switching induced by TGF- β (data not shown)²⁰.

In summary, our studies indicate that the LP of the small intestine contains B220⁺IgA⁺ cells that have just completed CSR. These cells are most probably generated *in situ* from LP B220⁺IgM⁺ B cells. LP-SCs facilitate IgA switching as well as differentiation into plasma cells with high efficiency. That antigen presentation might occur in the LP has been suggested by the finding that LP dendritic cells can directly sample intestinal pathogens²¹. Together, the interaction between LP B cells, dendritic cells and stromal cells would explain the T-cell-independent induction of IgA synthesis against commensal bacteria *in situ*. B1-cell-derived IgA²² is also specific for commensal bacteria and is generated independently of T cells or germinal centres²³. We propose that peritoneal B cells, which do not differentiate during migration through lymphoid organs²⁴⁻²⁶, will finally home to the gut LP, where, driven by the intestinal flora, they will switch and differentiate to IgA⁺ plasma cells. Thus, the gut LP may be not only an effector site but also an important inductive place for mucosal immune responses. This new perspective of the gut mucosa as a 'non-structured' lymphoid tissue offers an opportunity to improve the strategies for developing effective oral vaccines and mucosal adjuvants.

Note added in proof: Most LP and PP B cells expressed both

peripheral L-selectin and mucosal $\alpha 4\beta 7$ homing receptors, by surface staining. □

Methods

Mice

AID^{+/+}, AID^{+/-} and AID^{-/-} mice on a (CBA $\times$ C57BL/6) $\times$ C57BL/6 background were maintained in the animal facility of our department and used between 4 and 32 weeks of age. BALB/c and C57BL/6 mice were purchased from Japan SLC (Shizuoka, Japan) and used at 10–12 weeks. RAG-2^{-/-} mice were on a C57BL/6 background and used at 12–18 weeks of age. We maintained all mice in specific pathogen-free conditions.

Flow cytometric analysis

The following antibodies were used for staining: allophycocyanin (APC) or fluorescein isothiocyanate (FITC) anti-mouse B220, phycoerythrin (PE) anti-CD40, CD23, CD44, FITC-anti I-A^b (all from PharMingen); FITC or PE anti-IgM, IgA, FITC anti-IgG, Biotin anti-IgG3 (all from Southern Biotechnology Associates). PE-streptavidin was from Biomed. We performed cell analysis, multicolour flow-cytometry sorting of LP plasma cells, and cytospin preparations as described^{4,27}.

Preparation of cells and *in vitro* cell culture

For the time-course study of α CT production, IgA-depleted spleen cells from BALB/c mice were stained with FITC anti-IgA, incubated with anti-FITC-coated magnetic beads, and passed over a MACS column (Miltenyi Biotec). Cells from the negative fractions were incubated in RPMI medium supplemented with 10% fetal calf serum (FCS) and 50 μ M 2-mercaptoethanol (2-ME), stimulated with LPS, IL-5 and TGF- β for 5 days, washed, and further cultivated for 4 days without stimulation.

For *in vitro* IgA switching, we isolated LP and PP lymphocytes as described⁴. After staining with FITC anti-IgA, cells were incubated with anti-FITC, anti-CD43, anti-Thy-1.2 and anti-Mac-1-coated magnetic beads, and passed over a MACS column (Miltenyi Biotec). LP and PP B220⁺IgM⁺ cells were further sorted by FACS Vantage. Cells were cultivated and stimulated with LPS, IL-5 and TGF- β for 6 days. LP and PP B220⁺IgM⁺ cells from AID^{-/-} mice were positively selected with anti-B220-coated magnetic beads (Miltenyi Biotec), and activated with LPS (20 μ g ml⁻¹) for 48 h before retroviral transfection. LP-SCs from AID^{-/-} mice were isolated in LP lymphocytes prepared as described⁴, and co-purified with B cells from the B220-positive fractions, owing to their adherence to B cells. After 24 h of plating and removing nonadherent B cells, LP-SCs were incubated with RPMI medium supplemented with 10% FCS and 50 μ M 2-ME for 2 weeks to a semiconfluent culture, with half of the medium replaced every 3–4 days.

Macrophage and T-cell contamination in the LP-SC cultures was assessed by FACS using APC anti-Mac-1 and FITC anti-CD3 antibodies (PharMingen). No CD3⁺ cells could be detected and less than 3% of the cells expressed Mac-1. Spleen IgM⁺ B cells from AID^{+/+} mice purified by magnetic-bead depletion were cultured alone, co-cultured with LP-SCs or AC-11, or cultured with conditioned media (from a 4-day culture supernatant from LP-SCs from AID^{-/-} mice), in the presence of LPS, IL-5 and pan-specific anti-TGF- β antibodies (50 μ g ml⁻¹; R&D) where indicated. For all experiments, IgA switching stimulation was performed with recombinant mouse IL-5 (100 U ml⁻¹; R&D), human TGF- β 1 (1 ng ml⁻¹; R&D) and LPS (*Escherichia coli* O26:B6; 20 μ g ml⁻¹; Difco Laboratories), unless otherwise indicated.

Polymerase chain reaction with reverse transcription

LP B220⁺IgM⁺, B220⁺IgA⁺ and B220⁺IgA⁺ cells and PP B220⁺IgM⁺ and B220⁺IgA⁺ cells were isolated from BALB/c mice and sorted by FACS Vantage. CH12F3 cells were stimulated as described⁹. Total RNA was extracted using TRIzol (Gibco BRL), according to the manufacturer's instructions. Oligo(dT)-primed complementary DNA was prepared by reverse transcriptase. AID, α GT and I μ -C α transcripts were amplified by 30 cycles of PCR (except for AID transcripts in *in vitro* cultures, which were amplified by 35 cycles) using described primers³. α CT transcripts were amplified by nested PCR (30 cycles each), to obtain 317- and 171-base-pair products, using the following primer pairs: I α 253F, 5'-CCAGGCATGGTTGAGATAGATAG-3'; C μ 94R, 5'-AATGGTGCT GGGCAG-GAAGT-3'; I α 254F, 5'-CAGGCATGGTTGAGATAGATAG-3'; C μ 93R, 5'-ATGGTGCTGGGCAGGAAGT-3'. IL-6, IL-10 and TGF- β transcripts were amplified by 35 cycles of PCR with primers from Clontech.

Transfer experiments

We injected LP, PP and MLN IgA-depleted cells from C57BL/6 mice (1.5 $\times$ 10⁵ B220⁺IgM⁺ cells) into RAG-2^{-/-} mice through the retro-orbital plexus. After injection, mice were maintained on aqueous antibiotics for the duration of the experiment, as described²⁸. IgA levels in the faecal extracts of transferred mice were determined by enzyme-linked immunosorbent assay (ELISA) as described⁴.

Recombinant retrovirus production and infection of B cells

Mutant AID (AIDm-1) was generated by deleting the region between the *PmaCI* restriction sites in AID cDNA as described⁶. This deleted AID has lost most of the cytidine deaminase motif. We subcloned AID or AIDm-1 into pMX-IRES-GFP, which was constructed as described²⁹. The plasmids were transfected into the packaging cell line plat-E³⁰ by FuGENE reagent (Roche). Preactivated cells were suspended at a density of 4 $\times$ 10⁵ cells per ml in the medium containing retrovirus supplemented with 16 μ g ml⁻¹

polybrene (Sigma), stimulated for IgA switching, centrifuged at 3,000 r.p.m. for 45 min at 32 °C, and incubated for 48 h. The cells were then washed and incubated for another 3 days under the same stimulation conditions with or without LP-SCs.

Received 30 May; accepted 20 July 2001.

- Craig, S. W. & Cebra, J. J. Peyer's patches: an enriched source of precursors for IgA-producing immunocytes in the rabbit. *J. Exp. Med.* **134**, 188–200 (1971).
- Tseng, J. Transfer of lymphocytes of Peyer's patches between immunoglobulin allotype congenic mice: repopulation of the IgA plasma cells in the gut lamina propria. *J. Immunol.* **127**, 2039–2043 (1981).
- Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).
- Kamata, T. *et al.* Increased frequency of surface IgA-positive plasma cells in the intestinal lamina propria and decreased IgA excretion in hyper IgA (HIGA) mice, a murine model of IgA nephropathy with hypersecreted IgA. *J. Immunol.* **165**, 1387–1394 (2000).
- Lebman, D. A., Griffin, P. M. & Cebra, J. J. Relationship between expression of IgA by Peyer's patch cells and functional IgA memory cells. *J. Exp. Med.* **166**, 1405–1418 (1987).
- Muramatsu, M. *et al.* Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J. Biol. Chem.* **274**, 18470–18476 (1999).
- Stavnezer-Nordgren, J. & Sirlin, S. Specificity of immunoglobulin heavy chain switch correlates with activity of germline heavy chain genes prior to switching. *EMBO J.* **5**, 95–102 (1986).
- Iwasato, T., Shimizu, A., Honjo, T. & Yamagishi, H. Circular DNA is excised by immunoglobulin class switch recombination. *Cell* **62**, 143–149 (1990).
- Nakamura, M. *et al.* High frequency class switching of an IgM⁺B lymphoma clone CH12F3 to IgA⁺ cells. *Int. Immunol.* **8**, 193–201 (1996).
- Waldschmidt, T. J., Conrad, D. H. & Lynch, R. G. The expression of B cell surface receptors. I. The ontogeny and distribution of the murine B cell IgE Fc receptor. *J. Immunol.* **140**, 2148–2154 (1988).
- Waldschmidt, T. J., Kroese, F. G., Tytgret, L. T., Conrad, D. H. & Lynch, R. G. The expression of B cell surface receptors. III. The murine low-affinity IgE Fc receptor is not expressed on Ly 1 or 'Ly 1-like' B cells. *Int. Immunol.* **3**, 305–315 (1991).
- Oliver, A. M., Martin, F., Gartland, G. L., Carter, R. H. & Kearney, J. F. Marginal zone B cells exhibit unique activation, proliferative and immunoglobulin secretory responses. *Eur. J. Immunol.* **27**, 2366–2374 (1997).
- Martin, F. & Kearney, J. F. CD21^{high}IgM^{high} splenic B cells enriched in the marginal zone: distinct phenotypes and functions. *Curr. Top. Microbiol. Immunol.* **246**, 45–50 (1999).
- Whitlock, C. A., Tidmarsh, G. F., Muller-Sieburg, C. & Weissman, I. L. Bone marrow stromal cell lines with lymphopoietic activity express high levels of a pre-B neoplasia-associated molecule. *Cell* **48**, 1009–1021 (1987).
- Xu, J. *et al.* Mice deficient for the CD40 ligand. *Immunity* **1**, 423–431 (1994).
- Renshaw, B. R. *et al.* Humoral immune responses in CD40 ligand-deficient mice. *J. Exp. Med.* **180**, 1889–1900 (1994).
- Coffman, R. L., Lebman, D. A. & Shrader, B. Transforming growth factor β specifically enhances IgA production by lipopolysaccharide-stimulated murine B lymphocytes. *J. Exp. Med.* **170**, 1039–1044 (1989).
- Sonoda, E. *et al.* Transforming growth factor β induces IgA production and acts additively with interleukin 5 for IgA production. *J. Exp. Med.* **170**, 1415–1420 (1989).
- Ramsay, A. J. *et al.* The role of interleukin-6 in mucosal IgA antibody responses in vivo. *Science* **264**, 561–563 (1994).
- Shparago, N. *et al.* IL-10 selectively regulates murine Ig isotype switching. *Int. Immunol.* **8**, 781–790 (1996).
- Rescigno, M. *et al.* Dendritic cells express tight junction proteins and penetrate gut epithelial monolayers to sample bacteria. *Nature Immunol.* **2**, 361–367 (2001).
- Kroese, F. G. *et al.* Many of the IgA producing plasma cells in murine gut are derived from self-replenishing precursors in the peritoneal cavity. *Int. Immunol.* **1**, 75–84 (1989).
- Macpherson, A. J. *et al.* A primitive T cell-independent mechanism of intestinal mucosal IgA responses to commensal bacteria. *Science* **288**, 2222–2226 (2000).
- Fagarasan, S., Watanabe, N. & Honjo, T. Generation, expansion, migration and activation of mouse B1 cells. *Immunol. Rev.* **176**, 205–215 (2000).
- Fagarasan, S. & Honjo, T. T-independent immune response: new aspects of B cell biology. *Science* **290**, 89–92 (2000).
- Watanabe, N. *et al.* Migration and differentiation of autoreactive B-1 cells induced by activated γ/δ T cells in antierthyroid immunoglobulin transgenic mice. *J. Exp. Med.* **192**, 1577–1586 (2000).
- Fagarasan, S. *et al.* Alymphoplasia (aly)-type nuclear factor κ B-inducing kinase (NIK) causes defects in secondary lymphoid tissue chemokine receptor signaling and homing of peritoneal cells to the gut-associated lymphatic tissue system. *J. Exp. Med.* **191**, 1477–1486 (2000).
- Ikuta, K. *et al.* A developmental switch in thymic lymphocyte maturation potential occurs at the level of hematopoietic stem cells. *Cell* **62**, 863–874 (1990).
- Ye, S. K. *et al.* Induction of germline transcription in the TCR γ locus by Stat5: implications for accessibility control by the IL-7 receptor. *Immunity* **11**, 213–223 (1999).
- Morita, S., Kojima, T. & Kitamura, T. Plat-E: an efficient and stable system for transient packaging of retroviruses. *Gene Ther.* **7**, 1063–1066 (2000).
- Kinoshita, K., Harigai, M., Fagarasan, S., Muramatsu, M. & Honjo, T. A hallmark of active class switch recombination: Transcripts directed by I promoters on looped-out circular DNAs. *Proc. Natl. Acad. Sci. USA* (in the press).

Acknowledgements

We are especially grateful to M. Tanaka for cell sorting, and to Y. Tabuchi, T. Toyoshima and E. Inoue for their technical assistance. This work was supported by Center of Excellence Grant from the Ministry of Education, Science, Sports and Culture of Japan.

Correspondence and requests for materials should be addressed to T.H. (e-mail: honjo@mfour.med.kyoto-u.ac.jp).

Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly

Ming-Juan Luo*, Zhaolan Zhou*, Keith Magni*, Claudine Christoforides*, Juri Rappsilber†, Matthias Mann† & Robin Reed*

* Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA

† Protein Interaction Laboratory in the Center of Experimental Bioinformatics, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark

Recent studies indicate that splicing of pre-messenger RNA and export of mRNA are normally coupled *in vivo*^{1–6}. During splicing, the conserved mRNA export factor Aly is recruited to the spliced mRNA–protein complex (mRNP), which targets the mRNA for export. At present, it is not known how Aly is recruited to the spliced mRNP. Here we show that the conserved DEAD-box helicase UAP56, which functions during spliceosome assembly^{7–10}, interacts directly and highly specifically with Aly. Moreover, UAP56 is present together with Aly in the spliced mRNP. Significantly, excess UAP56 is a potent dominant negative inhibitor of mRNA export. Excess UAP56 also inhibits the recruitment of Aly to the spliced mRNP. Furthermore, a mutation in Aly that blocks its interaction with UAP56 prevents recruitment of Aly to

the spliced mRNP. These data suggest that the splicing factor UAP56 functions in coupling the splicing and export machineries by recruiting Aly to the spliced mRNP.

Splicing of pre-mRNA generates a specific mRNP complex that promotes mRNA export by recruiting the mRNA export factor Aly². Previous studies showed that a glutathione S-transferase (GST)–Aly fusion protein injected into oocytes of the frog *Xenopus* is specifically incorporated into this spliced mRNP². To screen for factors that mediate the recruitment of Aly to the spliced mRNP, we characterized the proteins that are bound to GST–Aly. Oocyte nuclei containing either GST–Aly or a GST control protein were incubated with glutathione beads and the bound proteins were analysed by SDS–polyacrylamide gel electrophoresis (SDS–PAGE) (Fig. 1a). Proteins unique to the Aly pull-down were micro-sequenced revealing the splicing factor UAP56 (ref. 7) as well as several highly abundant nuclear proteins (for example, nucleolin and lamin) (Fig. 1a, lane 2).

UAP56 is a member of the DEAD-box family of splicing factors, which have ATPase and RNA helicase motifs¹¹. The counterpart of UAP56 in the yeast *Saccharomyces cerevisiae*, Sub2p, has recently been identified and shown to function during spliceosome assembly^{8–10}. Aly also has a counterpart in *S. cerevisiae*, Yra1p, that is essential for mRNA export¹². The high levels of conservation between the yeast and metazoan splicing and export machineries suggest that the recruitment of Aly/Yra1p to the mRNP involves a conserved factor. Thus, our additional studies focused on the conserved splicing factor, UAP56.

To further investigate the relationship between Aly and UAP56, we asked whether these proteins also interact in mammalian cells. As shown by western analysis (Fig. 1b, top), Aly was specifically

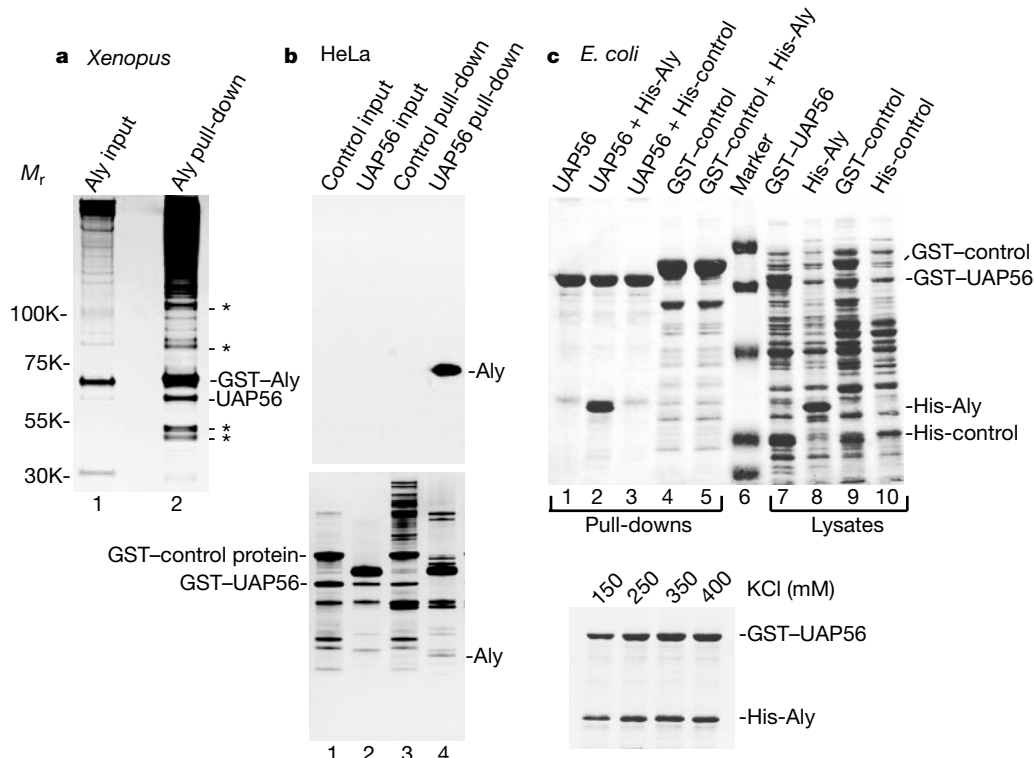


Figure 1 The mRNA export factor Aly interacts with the splicing factor UAP56. Aly interacts with UAP56 in *Xenopus* oocyte nuclei (a), HeLa nuclear extract (b) and *E. coli* lysate (c). a, GST–Aly (400 fmol) (lane 1) was injected into oocyte cytoplasm for import into nuclei followed by binding to glutathione beads². Bound proteins were detected on an SDS gel by silver stain. Mass spectrometry identified UAP56 and proteins (from top, indicated by asterisks: nucleolin, lamin L, TAF-1 β and nucleoplasmin). b, GST control

protein or GST–UAP56 (1 μ g each, lanes 1, 2) was incubated with HeLa nuclear extract (7.5 μ l). Proteins bound to glutathione beads (lanes 3 and 4) were analysed by western blot (top) or silver staining (bottom). c, *E. coli* lysates containing GST–UAP56, His–Aly or negative control proteins were mixed together in binding buffer containing KCl at 250 mM (top) or the indicated concentrations (bottom). Proteins bound to glutathione beads and lysates were detected by Coomassie stain.

selected from HeLa nuclear extract by GST–UAP56, but not by a negative control GST-fusion protein. A silver-stained gel containing the same samples showed that the pull-down of Aly by GST–UAP56 is highly specific as few bands other than Aly were detected (Fig. 1b, bottom, lane 4; the band designated Aly was confirmed by micro-sequencing). We conclude that Aly and UAP56 can associate with each other in both *Xenopus* and HeLa nuclei.

To determine whether Aly and UAP56 interact directly, GST–UAP56 and histidine-tagged Aly (His–Aly) were expressed in *Escherichia coli*, the total lysates containing each protein were mixed, and glutathione beads were added. A stoichiometric complex is formed between UAP56 and Aly (Fig. 1c, top, lane 2). The interaction between these two proteins is specific, as GST–UAP56 does not interact with any other protein in the total *E. coli* lysates. Moreover, no interaction was detected between GST–UAP56 and a negative control His-fusion protein (lane 3) nor between His–Aly and a negative control GST-fusion protein (lane 5). The interaction between Aly and UAP56 is also resistant to RNase (data not shown). We conclude that Aly and UAP56 interact directly and with high specificity. Moreover, the complex formed between these two proteins is stable in salt concentrations ranging from 150 to 400 mM, indicating that Aly and UAP56 interact tightly (Fig. 1c, bottom).

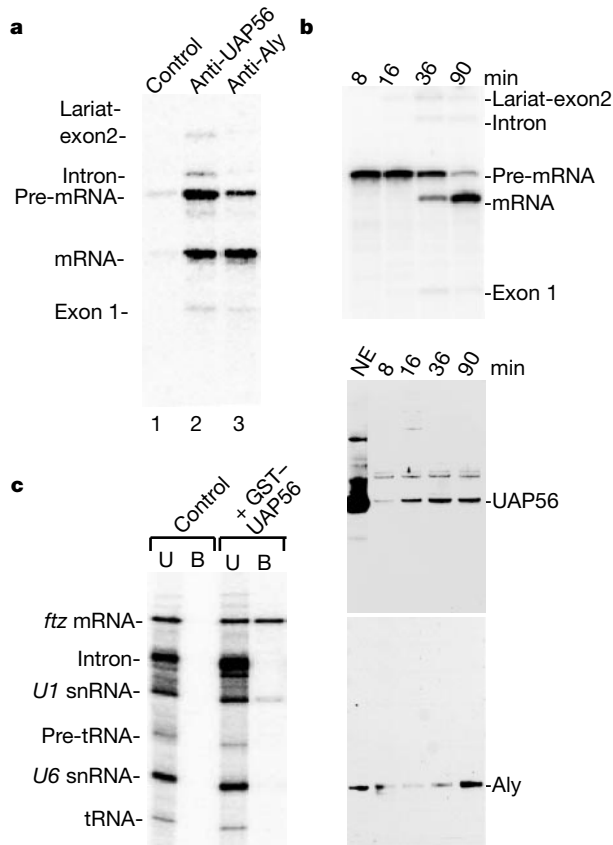


Figure 2 UAP56 is present in the spliced mRNP. **a**, *AdML* pre-mRNA was incubated in HeLa nuclear extract under splicing conditions for 45 min. After gel filtration, the fraction containing spliceosomes and spliced mRNP was incubated with protein A beads coupled to control, UAP56 or Aly antibodies. **b**, Spliceosomal complexes assembled on *AdML* pre-mRNA for the indicated times were purified^{2,20} in 60 mM NaCl. RNA (top panel) or western blot of purified complexes were probed with UAP56 (middle panel) or Aly (bottom panel) antibodies. NE, nuclear extract. **c**, Buffer or 55 fmol of GST–UAP56 was injected into oocyte cytoplasm. *ftz* pre-mRNA, pre-tRNA^{ser} and U1 and U6 snRNAs were then injected into oocyte nuclei and incubated for 1 h. GST pull-downs were carried out² and total unbound and bound RNAs analysed.

UAP56 was originally identified as a protein that interacts with U2AF65, a splicing factor that binds to the 3' splice site and recruits U2 snRNP (a small nuclear RNA–protein complex) early in spliceosome assembly⁷. Recent studies indicate that UAP56/Sub2p has multiple tasks in spliceosome assembly, including dissociation of U2AF65 from the spliceosome^{8–10}. Because UAP56 also interacts with an mRNA export factor (Aly), we asked whether UAP56, like Aly, is present in the spliced mRNP. Consistent with previous studies, an antibody to UAP56 immunoprecipitates the pre-mRNA (Fig. 2a, lane 2). Significantly, this antibody also immunoprecipitates the spliced mRNA, as does an antibody to Aly (Fig. 2a, lanes 2 and 3). These data suggest that UAP56 is present in the spliced mRNP. To verify that the immunoprecipitation of the spliced mRNP is not due to a crossreacting epitope, western analysis of purified spliceosomal complexes was carried out using the UAP56 and Aly antibodies. UAP56 is detected in the assembling spliceosome and is still present at 90 min, when the spliced mRNA is the major RNA species detected (Fig. 2b). Consistent with previous work², Aly is also detected at 90 min (Fig. 2b). Thus, the immunoprecipitation and western data indicate that UAP56 is present with Aly in the spliced mRNP.

To obtain further evidence that UAP56 is associated with the spliced mRNP, we carried out an *in vivo* pull-down assay. To do this, we injected GST–UAP56 into *Xenopus* oocytes and then injected *fushi tarazu* (*ftz*) pre-mRNA and control RNAs. After incubating the oocytes to allow splicing, nuclei were isolated and glutathione beads were added to the lysed nuclei. The spliced mRNA is efficiently bound to GST–UAP56 whereas the intron, U1 snRNA, pre-tRNA, U6 snRNA and tRNA are not (Fig. 2c). We conclude that UAP56, like Aly², associates with the spliced mRNP both *in vitro* and *in vivo*. Neither UAP56 nor Aly¹³ is detected in the spliced mRNP in the cytoplasm (data not shown).

UAP56 and Aly interact directly and both proteins are present in the spliced mRNP; this suggests that UAP56, in addition to functioning in splicing, functions with Aly in mRNA export. To test this possibility, we injected increasing amounts of His-tagged UAP56 into *Xenopus* oocytes and examined the export efficiency of several RNAs. Strikingly, excess UAP56 inhibits the export of *AdML* mRNA in a dose-dependent manner (Fig. 3a). The export of tRNA and U1 snRNA is not significantly affected by excess UAP56, indicating that the export block is specific to mRNA. The export inhibition seems to be general, as it was also observed with another mRNA (*ftz*, Fig. 3b). Analysis of the kinetics of export showed that excess UAP56 interferes with export throughout the time course (data not shown). We conclude that excess UAP56 functions as a dominant negative inhibitor of mRNA export. Notably, excess UAP56 does not affect export of the constitutive transport element (CTE) (Fig. 3c). This viral RNA element is exported by directly binding to Tip-associated protein (TAP), a late-acting mRNA export factor that interacts with Aly and the nuclear pore complex (reviewed in ref. 14). We conclude that excess UAP56 interferes with a factor or factors upstream of TAP in the mRNA export pathway.

Previous work showed that Aly is a limiting factor for mRNA export in oocytes (refs 2, 4; see also Fig. 4a, lanes 1, 2, 9–12). Thus, excess UAP56 may inhibit mRNA export by binding to and sequestering Aly, thereby preventing its recruitment to the spliced mRNP. Consistent with this possibility, the export block by excess UAP56 can be relieved by excess Aly in a dose-dependent manner (Fig. 4a, lanes 1–8). To determine whether excess UAP56 also prevents the recruitment of Aly to the spliced mRNP, we injected oocytes with GST–Aly without UAP56 or in the presence of increasing amounts of UAP56. Western analysis shows that GST–Aly is present at similar levels in all of the nuclei of injected oocytes (Fig. 4b, top panel). A GST pull-down assay was then used to detect Aly in the spliced mRNP. Excess UAP56 inhibits the recruitment of Aly to the spliced mRNP in a dose-dependent manner (Fig. 4b, bottom panel).

The data presented above provide evidence that UAP56 functions in recruiting Aly to the spliced mRNP. To further test this possibility, we generated a mutation in GST–Aly that inhibits its association with UAP56 (Fig. 5a). As shown by the *in vitro* binding assay (Fig. 5b), wild-type Aly or an Aly RGG-domain mutant (Δ RGG) binds UAP56 efficiently, whereas an Aly carboxy-terminal mutant (Δ C) does not. Both of the mutants are imported into oocyte nuclei (Fig. 5c) and bind TAP (data not shown) as efficiently as wild-type

Aly, indicating that the mutants are functional proteins.

To determine whether the Aly mutants can be incorporated into the spliced mRNP, we injected the mutant and wild-type proteins into oocytes and carried out a GST pull-down assay from the oocyte nuclei (Fig. 5d). Significantly, mRNA was efficiently pulled down by wild-type Aly and Δ RGG Aly but not by Δ C Aly. Thus, the Aly mutant that fails to bind UAP56 efficiently also inhibits the recruitment of Aly to the spliced mRNP. These data provide further

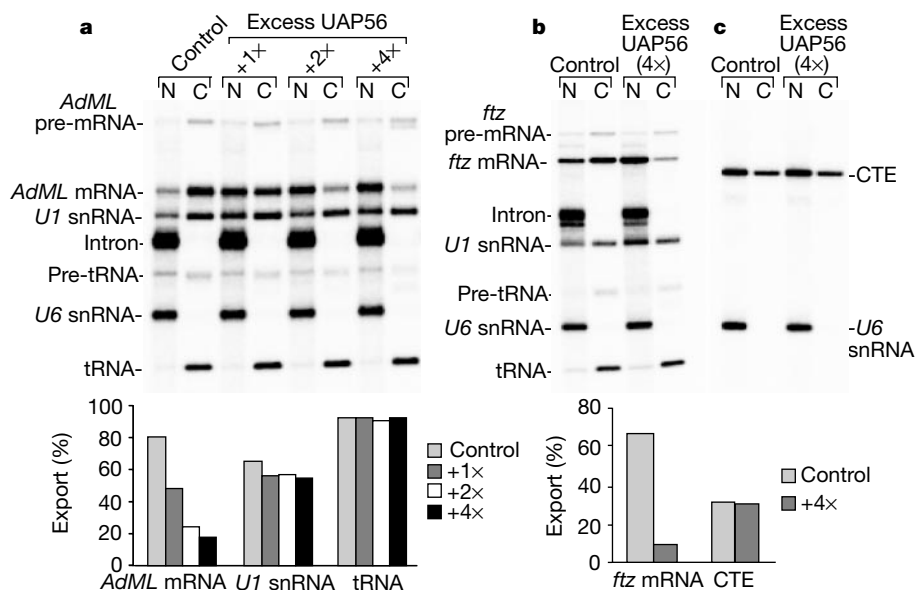


Figure 3 Excess UAP56 inhibits mRNA export. **a**, Buffer or His-UAP56 (1x = 140 fmol) was injected into oocyte cytoplasm. *AdML* pre-mRNA, pre-tRNA^{ser} and *U1* and *U6* snRNAs were then injected into nuclei. Oocytes were incubated for 3 h. **b**, Buffer or 4x His-UAP56 was injected as in **a**. *ftz* pre-mRNA, pre-tRNA^{ser} and *U1* and *U6* snRNAs were then injected

into nuclei and incubated for 2 h. **c**, Same as **b** except that CTE RNA and *U6* snRNA were injected into nuclei and incubated for 75 min. RNA from the nucleus (N) or cytoplasm (C) was analysed.

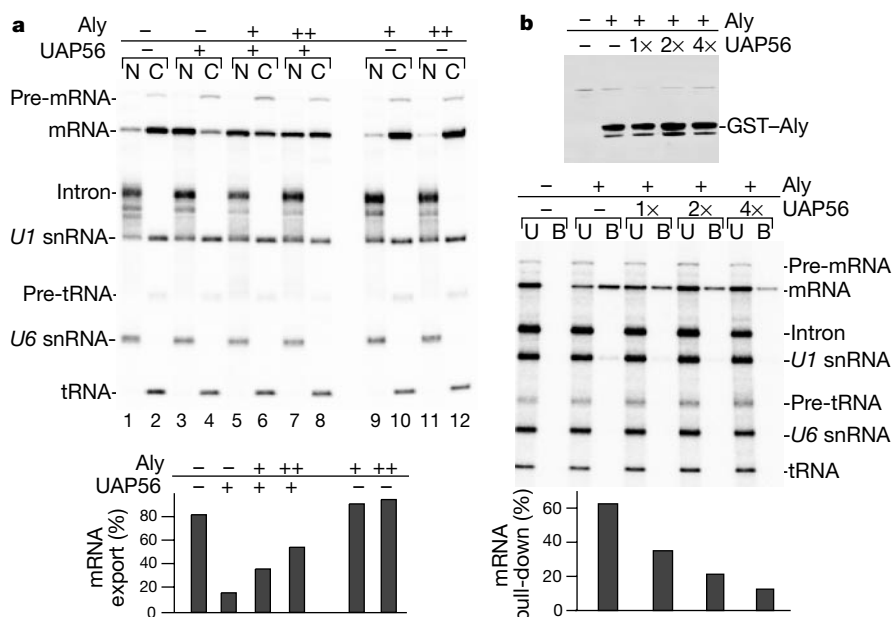


Figure 4 UAP56 recruits Aly to the spliced mRNP. **a**, Excess Aly relieves the UAP56 mRNA export block. Buffer (–), His-UAP56 (280 fmol, +), His-Aly (70 fmol, +) or His-Aly (140 fmol, ++), were injected into oocyte cytoplasm as indicated. *ftz* pre-mRNA, pre-tRNA^{ser} and *U1* and *U6* snRNAs were then injected into nuclei and incubated for 2.5 h. **b**, Excess UAP56 inhibits recruitment of Aly to the spliced mRNP. Buffer or His-UAP56 was

injected as in Fig. 3a. After overnight incubation, buffer or 50 fmol of GST–Aly was injected into cytoplasm and incubated for 4 h. Top, GST–Aly in nuclei was analysed by western blot. Bottom, *ftz* pre-mRNA, pre-tRNA^{ser} and *U1* and *U6* snRNAs were injected into nuclei and incubated for 45 min. GST pull-downs from nuclei were carried out.

evidence that UAP56 functions in recruiting Aly to the spliced mRNA.

In this study, we have shown that the mRNA export factor Aly, which is recruited to the mRNA during splicing, specifically interacts with the splicing factor UAP56. This protein is also present with Aly in the spliced mRNA, a complex that targets the mRNA for export. Functional studies show that excess UAP56 is a dominant negative inhibitor of mRNA export and prevents the recruitment of Aly to the spliced mRNA. In addition, a mutation in Aly that inhibits its binding to UAP56 also blocks recruitment of Aly to the spliced mRNA. Together, these data indicate that coupling between the splicing and export machineries involves a direct interaction between UAP56 and Aly. Our previous work indicated that the spliced mRNA is a large spliceosome-sized complex¹ and, thus, is likely to undergo significant remodelling before transit through the pore. The observation that UAP56 is present in the spliced mRNA only in the nucleus raises the possibility that this DEAD-box helicase functions in nuclear remodelling. A number of components of the mRNA export machinery are conserved from yeast to higher eukaryotes, including Aly, TAP/p15, hGle1, hGle2 and hDbp5p (reviewed in ref. 14). Recent studies revealed that the essential yeast splicing factor Sub2p is the homologue of UAP56 (refs 8–10). In an accompanying paper²³, Sub2p is shown to interact physically and genetically with the yeast homologue of Aly, Yra1p, and that Sub2p is required for export of mRNAs derived from both intron-containing and naturally intron-lacking genes. Thus,

UAP56/Sub2p may play a general conserved role in recruiting Aly/Yra1p to mRNPs. □

Methods

Plasmids encoding His-Aly (also named BEF)¹⁵, GST-Aly¹⁶, GST-UAP56⁷, *AdML* and *ftz* pre-mRNAs, *U1* and *U6* snRNAs and pre-tRNA^{Ser} were described^{1,2}. Negative control proteins were GST-Sir2¹⁷ and His-Ran¹⁸. The plasmid encoding His-UAP56 was constructed by inserting the cDNA of UAP56 into protein expression vector pET9d-HIS6-TB at *NcoI* and *BamHI* sites. The Aly Δ RGG mutant was obtained by polymerase chain reaction (PCR); the Δ C mutant was constructed by deletion of the *PstI* fragment that contains the Aly C-terminal coding region from the expression vector¹⁶. Cytoplasmic injection to import proteins², nuclear injection and oocyte isolation were carried out as described¹⁹. GST pull-down assay of the *in vivo* complexes was as previously described². Spliceosomal complexes were purified using affinity purification with maltose-binding protein²⁰. Western blots were probed with antibodies against UAP56⁷ and Aly²¹. Proteins that interact with GST-UAP56 or GST-Aly were identified by matrix-assisted laser desorption/ionization (MALDI) peptide mapping and nano-electrospray tandem mass spectrometry²².

Received 7 June; accepted 21 August 2001.

- Luo, M. & Reed, R. Splicing is required for rapid and efficient mRNA export in metazoans. *Proc. Natl Acad. Sci. USA* **96**, 14937–14942 (1999).
- Zhou, Z. *et al.* The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature* **407**, 401–405 (2000).
- Huang, Y. & Steitz, J. A. Splicing factors SRP20 and 9G8 promote the nucleocytoplasmic export of mRNA. *Mol. Cell* **7**, 899–905 (2001).
- Rodrigues, J. P. *et al.* REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus. *Proc. Natl Acad. Sci. USA* **98**, 1030–1035 (2001).
- Le Hir, H., Izaurralde, E., Maquat, L. E. & Moore, M. J. The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon–exon junctions. *EMBO J.* **19**, 6860–6869 (2000).
- Kataoka, N. *et al.* Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm. *Mol. Cell* **6**, 673–682 (2000).
- Fleckner, J., Zhang, M., Valcarcel, J. & Green, M. R. U2AF65 recruits a novel human DEAD box protein required for the U2 snRNP-branchpoint interaction. *Genes Dev.* **11**, 1864–1872 (1997).
- Kistler, A. L. & Guthrie, C. Deletion of MUD2, the yeast homolog of U2AF65, can bypass the requirement for sub2, an essential spliceosomal ATPase. *Genes Dev.* **15**, 42–49 (2001).
- Libri, D., Graziani, N., Saguez, C. & Boulay, J. Multiple roles for the yeast SUB2/yUAP56 gene in splicing. *Genes Dev.* **15**, 36–41 (2001).
- Zhang, M. & Green, M. R. Identification and characterization of yUAP/Sub2p, a yeast homolog of the essential human pre-mRNA splicing factor hUAP56. *Genes Dev.* **15**, 30–35 (2001).
- Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315–326 (1998).
- Strasser, K. & Hurt, E. Yra1p, a conserved nuclear RNA-binding protein, interacts directly with Mex67p and is required for mRNA export. *EMBO J.* **19**, 410–420 (2000).
- Kim, V. N. *et al.* The Y14 protein communicates to the cytoplasm the position of exon–exon junctions. *EMBO J.* **20**, 2062–2068 (2001).
- Conti, E. & Izaurralde, E. Nucleocytoplasmic transport enters the atomic age. *Curr. Opin. Cell Biol.* **13**, 310–319 (2001).
- Virbasius, C. M., Wagner, S. & Green, M. R. A human nuclear-localized chaperone that regulates dimerization, DNA binding, and transcriptional activity of bZIP proteins. *Mol. Cell* **4**, 219–228 (1999).
- Bruhn, L., Munnerlyn, A. & Grosschedl, R. ALY, a context-dependent coactivator of LEF-1 and AML-1, is required for TCR α enhancer function. *Genes Dev.* **11**, 640–653 (1997).
- Moazed, D., Kistler, A., Axelrod, A., Rine, J. & Johnson, A. D. Silent information regulator protein complexes in *Saccharomyces cerevisiae*: a SIR2/SIR4 complex and evidence for a regulatory domain in SIR4 that inhibits its interaction with SIR3. *Proc. Natl Acad. Sci. USA* **94**, 2186–2191 (1997).
- Izaurralde, E., Kutay, U., von Kobbe, C., Mattaj, I. W. & Gorlich, D. The asymmetric distribution of the constituents of the Ran system is essential for transport into and out of the nucleus. *EMBO J.* **16**, 6535–6547 (1997).
- Hamm, J. & Mattaj, I. W. Monomethylated cap structures facilitate RNA export from the nucleus. *Cell* **63**, 109–118 (1990).
- Das, R., Zhou, Z. & Reed, R. Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. *Mol. Cell* **5**, 779–787 (2000).
- Wichmann, I., Garcia-Lozano, J. R., Respalda, N., Gonzalez-Escribano, M. F. & Nunez-Roldan, A. Autoantibodies to transcriptional regulation proteins DEK and ALY in a patient with systemic lupus erythematosus. *Hum. Immunol.* **60**, 57–62 (1999).
- Vorm, O., Roepstorff, P. & Mann, M. Improved resolution and very high sensitivity in MALDI TDF of matrix surfaces made by fast evaporation. *Anal. Chem.* **66**, 3281–3287 (1994).
- Sträßer, K. & Hurt, E. Splicing factor Sub2p is required for nuclear mRNA export through its interaction with Yra1p. *Nature* **413**, 648–652 (1999).

Acknowledgements

We are grateful to M. Green for the UAP56 and His-BEF cDNAs and UAP56 antibody, to R. Grosschedl for the GST-Aly clone, to D. Moazed for the GST-Sir2 clone, and to A. Nunez-Roldan for the Aly antibody. We are indebted to D. Dorman, B. Lee and S. Lee for discussion and comments on the manuscript. We thank B. Lee for providing the GST-Aly Δ RGG protein. J.R. is a Marie Curie Fellow. M.M. was supported by a grant from the Danish National Research Foundation.

Correspondence and requests for materials should be addressed to R.R. (e-mail: rreed@hms.harvard.edu).

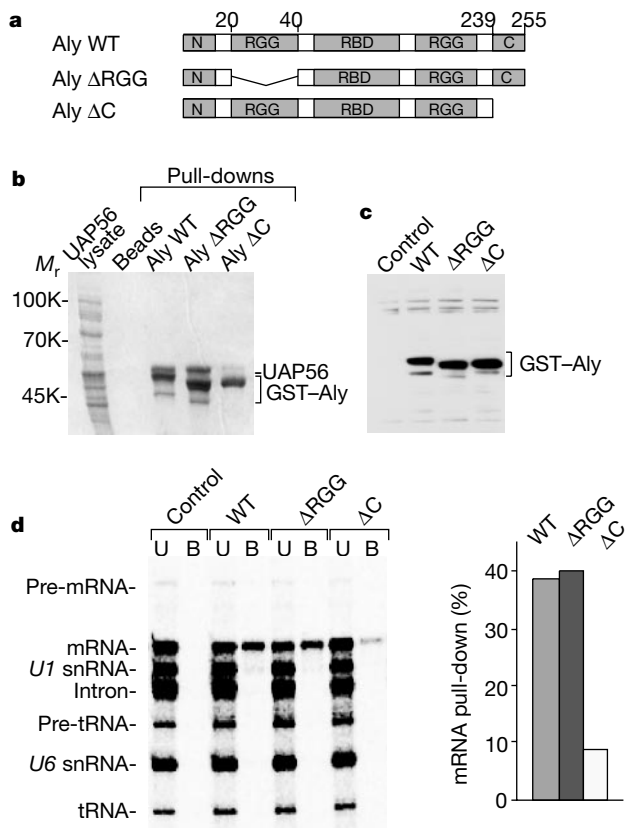


Figure 5 The UAP56–Aly interaction is required for Aly recruitment to the spliced mRNA. **a**, Schematic of Aly showing deletions. WT, wild type. **b**, GST–Aly wild type, Δ RGG or Δ C pre-bound to glutathione beads was rotated with His-UAP56 lysate. Bound proteins were analysed. **c**, Buffer or 50 fmol of GST–Aly wild type, Δ RGG or Δ C was injected into oocyte cytoplasm. GST–Aly in nuclei was analysed by western blot. **d**, Oocytes were injected as in **c**. *AdML* pre-mRNA, pre-tRNA^{Ser} and *U1* and *U6* snRNAs were then injected into nuclei and incubated for 1 h. GST pull-downs of nuclei were carried out.

Splicing factor Sub2p is required for nuclear mRNA export through its interaction with Yra1p

Katja Sträßer & Ed Hurt

BZH—Biochemie-Zentrum Heidelberg, Im Neuenheimer Feld 328, D-69120 Heidelberg, Germany

The yeast nuclear protein Yra1p is an essential export factor for mRNA. Yra1p interacts directly with the mRNA transport factor Mex67p/Mtr2p, which is associated with the nuclear pore^{1,2}. Here, we report a genetic interaction between *YRA1* and *SUB2*, the gene for a DEAD box helicase involved in splicing^{3–5}. Mutation of *SUB2* as well as its overexpression leads to a defect in mRNA export. Moreover, Yra1p and Sub2p bind directly to each other both *in vivo* and *in vitro*. Significantly, Sub2p and Mex67p/Mtr2p bind to the same domains of Yra1p, and the proteins compete for binding to Yra1p. Together, these data indicate that the spliceosomal component Sub2p is also important in mRNA export and may function to recruit Yra1p to the mRNA. Sub2p may then be displaced from Yra1p by the binding of Mex67p/Mtr2p, which participates in the export of mRNA through the nuclear pores.

The essential yeast mRNA export factors Yra1p and Mex67p/Mtr2p bind to each other *in vivo* as well as *in vitro*². Despite their interaction, Yra1p localizes to the nucleus^{2,6}, whereas the Mex67p/Mtr2p complex is enriched at the nuclear pores¹. Thus, the major pools of these two mRNA export factors are separated *in vivo*. To identify potential interaction partners of Yra1p, which could be components of the intranuclear mRNA export machinery, we sought to perform a synthetic lethal screen with mutant alleles of *YRA1*.

To find an appropriate mutant allele for this approach, we carried out a functional analysis of the domains in Yra1p. The highly conserved N1 domain, the variable, RG-rich N2 domain, the RNA recognition (RRM) domain, and the basic C1 domain of Yra1p are indicated in Fig. 1a. Surprisingly, deletion of any one of these domains does not lead to a growth defect (Fig. 1b, top rows; see Fig. 1a for deleted amino acids). However, *yra1-ΔN1* and *yra1-ΔRRM* are synthetically lethal with a mutant allele of *MEX67*, *mex67-5* (Fig. 1b, bottom rows).

In light of this observation and the possibility that the *yra1-ΔRRM* allele may identify RNA-binding proteins, we used this allele for the synthetic lethal screen. In this screen, 30 synthetic lethal mutants were isolated. The wild-type gene that rescued the synthetic lethality of one of the mutants was cloned by complementation and shown to be *SUB2* (data not shown). Interestingly, Sub2p was recently identified as an essential pre-mRNA splicing factor critical in spliceosome assembly^{3–5}. This protein and its mammalian homologue UAP56 are members of the DEAD box family of RNA helicases^{3–5}. From the remaining 29 synthetically lethal mutants, 13 were complemented by *SUB2*, 6 by *MEX67* and 2 by *MTR2*, indicating that the *yra1-ΔRRM* synthetic lethal screen predominantly yields factors involved in mRNA export, and thus is highly specific for this export pathway.

To test for synthetic lethality between *SUB2* and *YRA1* directly, mutant alleles of *yra1* and *sub2* were combined. As shown in Fig. 1c, *yra1-ΔRRM* and *yra1-ΔN1* are synthetically lethal with a *sub2* mutant allele, *sub2-85* (see below). This finding confirms the synthetically lethal relationship found in the screen. In contrast, *sub2-85* is not synthetically lethal with mutants of *MEX67* (data not shown). This result demonstrates a tight genetic interaction between *YRA1* and *SUB2*.

The genetic interaction between *SUB2* and the mRNA factor

YRA1 prompted us to test whether Sub2p functions not only in splicing, but also in mRNA export. To this end, we monitored nuclear export of poly(A)⁺ RNA in yeast cells depleted of the essential Sub2p. Expression of a protein A (ProtA)-tagged version of Sub2p under control of the *GAL1* promoter, but not under the control of the constitutive *NOPI* promoter, was terminated by shifting the cells from galactose- to glucose-containing medium (Fig. 2a, left panel). Significantly, cells depleted of Sub2p greatly accumulate mRNA in the nucleus (Fig. 2a, right panel). Furthermore, overexpression of Sub2p by the strong *GAL1* promoter (Fig. 2a, left panel) also causes a nuclear accumulation of poly(A)⁺ RNA (Fig. 2a, right panel). The latter result suggests that excess Sub2p impairs mRNA export in a dominant negative manner by titrating one or more other mRNA export factors.

To verify a role for Sub2p in mRNA export, we generated a thermosensitive mutant of *SUB2* (Fig. 2b, left panel). Analysis of poly(A)⁺ RNA localization in *sub2-85* thermosensitive cells revealed normal export of mRNA at the permissive temperature but a significant accumulation inside the nucleus shortly (15 min) after

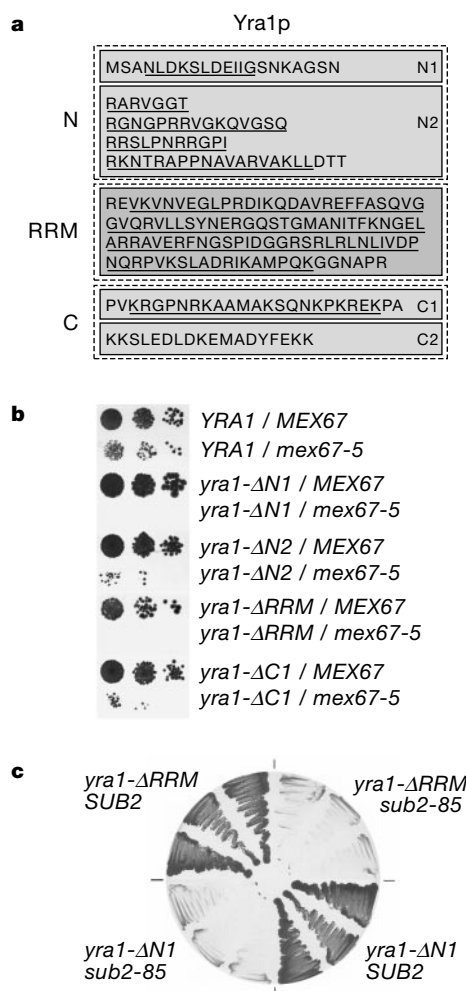


Figure 1 *YRA1* interacts genetically with *MEX67* and *SUB2*. **a**, Domain organization of Yra1p. Deletions are underlined. All deletion mutants are viable in a *yra1Δ* as well as a *yra1Δ/yra2Δ* background. Yra2p (YKL214c) is homologous to Yra1p, and *YRA2* on an *ARS/CEN* plasmid rescues the nonviable *yra1* null mutant (data not shown). **b**, Synthetic lethality of *yra1-ΔN1* and *yra1-ΔRRM* with *mex67-5* assessed on 5-fluoro-orotic acid (5-FOA) plates (5 d at 30 °C). **c**, Synthetic lethality of *yra1-ΔN1* and *yra1-ΔRRM* with *sub2-85*. A *yra1/sub2* double-shuffle strain containing the indicated alleles of *YRA1* and *SUB2* was grown on a 5-FOA plate (5 d at 23 °C).

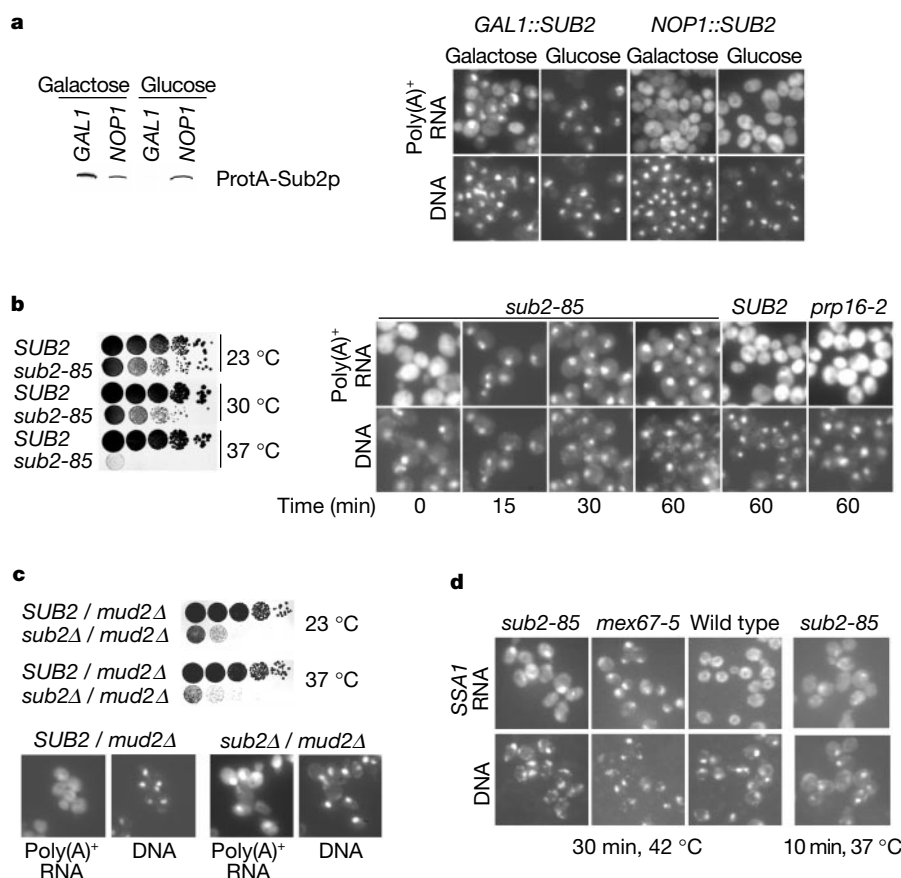


Figure 2 Sub2p is involved in nuclear mRNA export. **a**, Depletion or overexpression of Sub2p leads to an mRNA export defect. Left, western blot showing Sub2p expression. Right, localization of poly(A)⁺ RNA in GAL1::SUB2 cells grown in galactose (overexpression) or shifted to glucose medium (10 h depletion). **b**, Left, growth of SUB2 or sub2-85 thermosensitive cells at different temperatures. Right, mRNA accumulation in

SUB2, sub2-85 and prp16-2 cells shifted from 23 to 37 °C for the indicated times. **c**, Growth of the sub2Δ/mud2Δ and mud2Δ strains (top) and localization of poly(A)⁺ RNA (bottom). **d**, sub2-85 and mex67-5 cells were shifted for the indicated periods to the indicated temperatures, and the SSA1 mRNA (intron-lacking transcript) was localized by *in situ* hybridization.

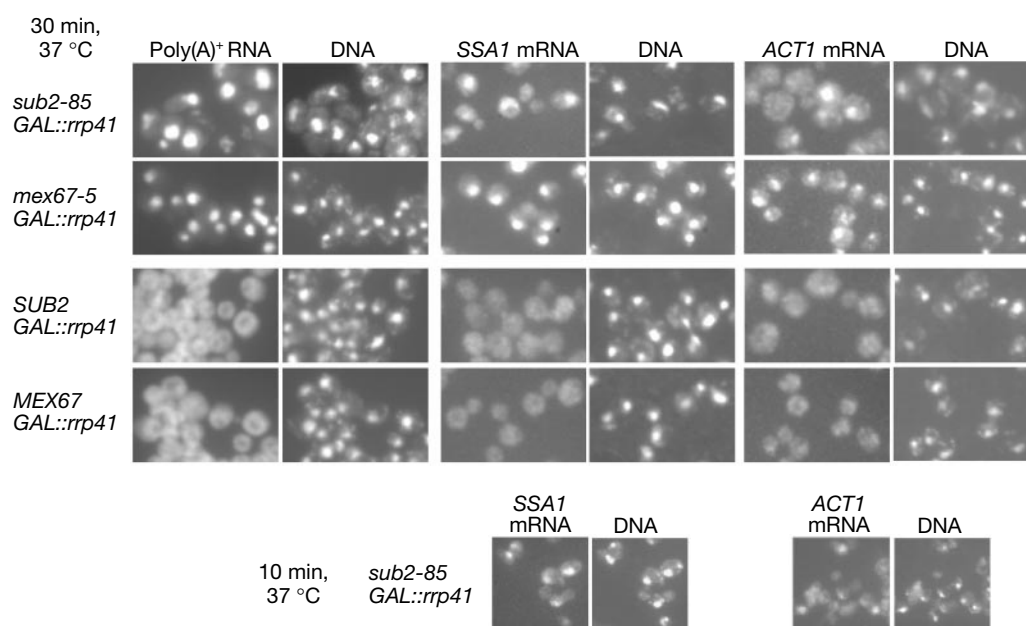


Figure 3 Sub2p is involved in nuclear mRNA export of both intron-containing and intron-lacking mRNAs. sub2-85/GAL::rrp41, mex67-5/GAL::rrp41, SUB2/GAL::rrp41 and MEX67/GAL::rrp41 strains were depleted for Rrp41p (10 h in glucose-containing

medium) and shifted for 10 or 30 min to 37 °C. SSA1 and ACT1 mRNAs as well as poly(A)⁺ RNAs were localized by *in situ* hybridization.

shifting to the restrictive temperature (Fig. 2b, right panel). The fast onset of nuclear accumulation indicates a direct role for Sub2p in mRNA export. In contrast to the *sub2-85* mutant, thermosensitive splicing mutants such as *prp16-1* (Fig. 2b, right panel) or *prp4-2* (data not shown) do not show a significant mRNA export defect. Another nuclear RNA export pathway, transfer RNA export, is not impaired in the *sub2-85* thermosensitive mutant (data not shown). These latter two results, together with the observations that *in vivo* depletion of Sub2p, overexpression of Sub2p, and a thermosensitive mutant of Sub2p all result in nuclear accumulation of mRNA, demonstrate that the splicing factor Sub2p also has a specific function in the export of mRNA.

It was recently shown that deletion of *MUD2* (another splicing factor) can bypass the requirement for Sub2p (ref. 5). Consistent with this study, we found that a *sub2Δ/mud2Δ* double mutant is viable, but grows extremely slowly at physiological temperatures (for example, at 23 and 30 °C) (Fig. 2c, top panel). These data indicate that Sub2p has functions other than releasing Mud2p from the spliceosome. Consistent with the slow-growing phenotype, poly(A)⁺ RNA export is still impaired in the *sub2Δ/mud2Δ* double mutant, although the degree of nuclear mRNA accumulation is slightly less than in the *sub2* thermosensitive mutant (Fig. 2c, bottom panel). We conclude that Sub2p is an important mRNA export factor, the function of which is only partly restored through deletion of *MUD2*.

We next investigated whether Sub2p is preferentially involved in the nuclear export of intron-containing mRNAs, which require

splicing before export, or whether it also participates in the export of mRNAs naturally lacking introns. As an example of an mRNA naturally lacking introns, we used the heat shock mRNA (*SSA1*)^{7,8}. Notably, nuclear export of *SSA1* mRNA is not only impaired in the *mex67-5* mutant (see also ref. 7), but also in the *sub2-85* thermosensitive mutant (Fig. 2d). We conclude that Sub2p is important in the export of mRNAs lacking introns.

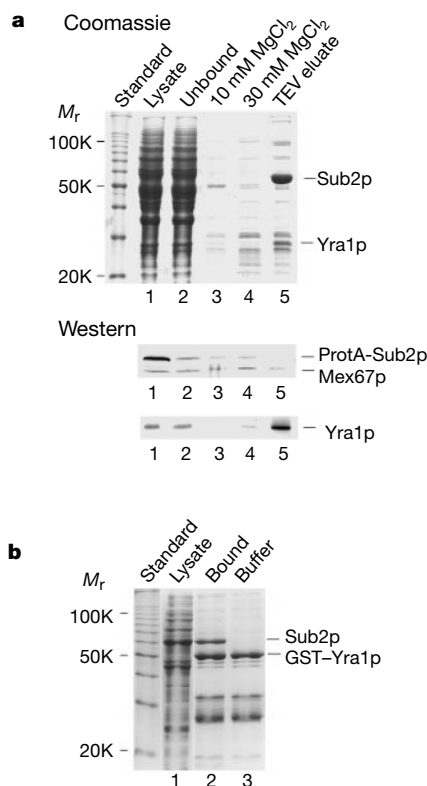


Figure 4 Yra1p interacts with Sub2p *in vivo* and *in vitro*. **a**, ProtA-TEV-Sub2p was affinity purified on IgG Sepharose. Total cell lysate (lane 1), unbound fraction (lane 2), washes with 10 mM MgCl₂ (lane 3) and 30 mM MgCl₂ (lane 4), and elution with TEV protease (lane 5) were analysed by SDS-PAGE. Coomassie staining (top) and western blotting with the indicated antibodies (bottom) are shown. **b**, *E. coli* lysate containing Sub2p (lane 1) or buffer was incubated with GST-Yra1p. Bound fractions were analysed by SDS-PAGE and Coomassie staining.

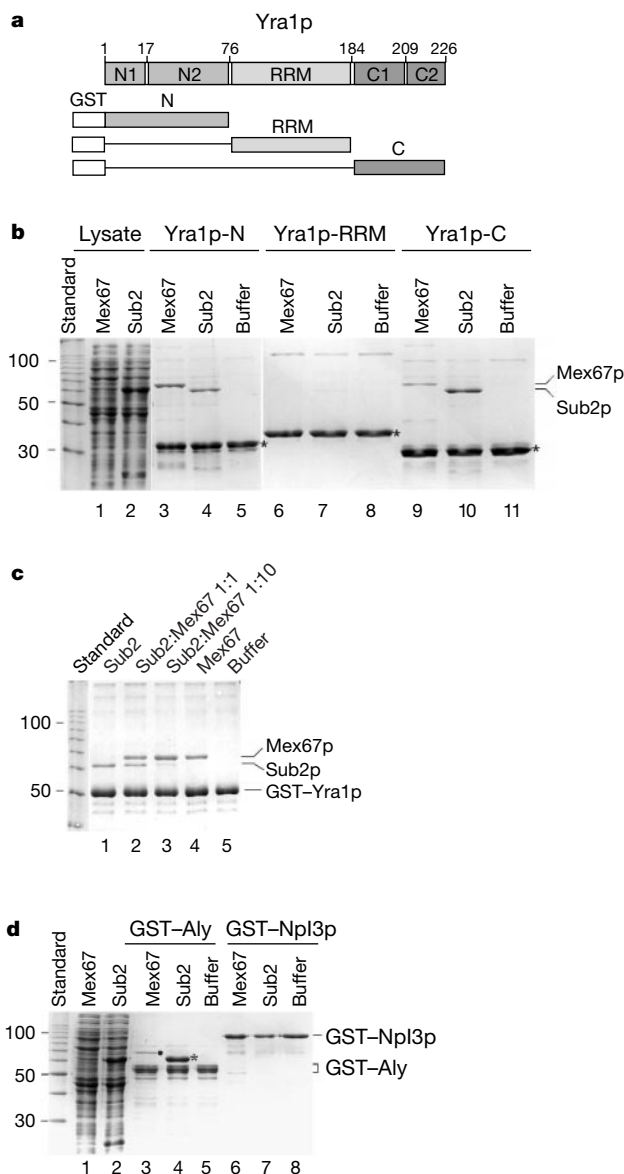


Figure 5 Mex67p/Mtr2p competes with Sub2p for binding to Yra1p. **a**, Domain organization of Yra1p and GST-fusion proteins used for *in vitro* binding. **b**, GST-N-domain (lanes 3–5), GST-RRM-domain (lanes 6–8), or GST-C-domain (lanes 9–11) of Yra1p immobilized on glutathione sepharose beads were incubated with *E. coli* lysate containing Mex67p/Mtr2p (lane 1, input; lanes 3, 6, 9, pull-downs), Sub2p (lane 2, input; lanes 4, 7, 10, pull-downs), or buffer (lanes 5, 8, 11, pull-downs). Asterisks indicate the position of the GST-Yra1p domains. **c**, GST-Yra1p was incubated with *E. coli* lysates containing Sub2p (lane 1), Sub2p plus equimolar (lane 2) or tenfold molar excess of Mex67p/Mtr2p (lane 3), Mex67p/Mtr2p only (lane 4), or buffer only (lane 5). **d**, Human Aly binds to Mex67p/Mtr2p (filled circle) and Sub2p (asterisk) *in vitro*. GST-Aly (lanes 3–5) or GST-Npl3p (lanes 6–8) were incubated with *E. coli* lysate containing Mex67p/Mtr2p (lane 1, input; lanes 3, 6, pull-downs), Sub2p (lane 2, input; lanes 4, 7, pull-downs) or buffer (lanes 5, 8, pull-downs).

We could not detect nuclear accumulation of an intron-containing mRNA (for example, actin mRNA) using *in situ* hybridization, in either the *mex67-5* or the *sub2-85* thermosensitive mutant (data not shown). This may be due to the lower abundance of actin transcripts compared with *SSA1* mRNAs, which thus remain under the *in situ* detection limit. To assess whether Sub2p and Mex67p are involved in the nuclear export of intron-containing mRNAs, we sought to stabilize intranuclear actin mRNA by slowing down pre-mRNA degradation. Previous studies showed that nuclear turnover of pre-mRNA is delayed in the exosomal *GAL::RRP41* mutant⁹. We therefore repressed expression of the *RRP41* gene (by growth in glucose-containing medium) in the *mex67-5* and *sub2-85* thermosensitive mutants, before shifting to the restrictive temperature. Importantly, wild-type *MEX67* or *SUB2* cells with repressed *GAL::RRP41* do not exhibit nuclear accumulation of poly(A)⁺, *SSA1* or *ACT1* mRNAs at 37°C (Fig. 3). In contrast, *ACT1* mRNA, *SSA1* mRNA and poly(A)⁺ RNA accumulate in the nucleus of *mex67-5* and *sub2-85* thermosensitive cells on shifting to the restrictive temperature with repressed *RRP41* (Fig. 3). We conclude that Mex67p and Sub2p are involved in the nuclear export of both intron-containing and intron-lacking mRNA.

That *YRA1* and *SUB2* interact genetically and are both essential for mRNA export prompted us to look for a physical interaction between these two proteins. First, we tested whether Yra1p is in a complex with Sub2p *in vivo*. We affinity purified functional ProtA-tagged Sub2p from yeast. As shown in Fig. 4a, Sub2p is efficiently purified from yeast cells together with Yra1p. In contrast, Mex67p is not enriched in the eluate (Fig. 4a, Mex67p, compare lanes 1 and 5). We conclude that Yra1p specifically interacts with Sub2p *in vivo*.

We next tested whether the interaction between Sub2p and Yra1p is direct. Glutathione S-transferase (GST)-tagged full-length Yra1p was immobilized on beads and incubated with an *Escherichia coli* lysate containing Sub2p. As shown in Fig. 4b (lane 2), Sub2p binds to GST–Yra1p, whereas the other *E. coli* proteins in the lysate do not. Furthermore, this interaction between Sub2p and Yra1p is not RNA dependent, as the *E. coli* lysate was incubated with RNase before being added to GST–Yra1p. Thus, Yra1p and Sub2p interact directly both *in vivo* and *in vitro*.

Previous studies showed that Yra1p also binds directly to the Mex67p/Mtr2p complex^{2,10}. Therefore, we sought to identify the domains of Yra1p that interact with Sub2p or Mex67p/Mtr2p. We expressed GST-fusion proteins containing the N domain, the RRM domain, or the C domain of Yra1p (see Fig. 5a) and tested them for binding to either the Mex67p/Mtr2p complex or Sub2p. Significantly, both Mex67p/Mtr2p and Sub2p bind to the N domain (Fig. 5b, lanes 3 and 4) and the C domain of Yra1p (Fig. 5b, lanes 9 and 10). In contrast, the RRM domain does not bind to either protein (Fig. 5b, lanes 6 and 7). We conclude that Sub2p and Mex67p/Mtr2p bind to the same regions of Yra1p.

That Sub2p and Mex67p interact with the same regions of Yra1p raised the possibility that there is a temporal order of binding to Yra1p. Specifically, Sub2p may bind to Yra1p first and then be replaced by Mex67p. To test this possibility, we carried out a competition assay. An *E. coli* lysate containing Sub2p was mixed with an equimolar amount or a tenfold excess of purified Mex67p/Mtr2p complex and added to GST–Yra1p on beads. Mex67p competed with Sub2p for binding to Yra1p when Mex67p and Sub2p were present in equimolar amounts (Fig. 5c, lane 2). When Mex67p is present in a tenfold excess over Sub2p, Mex67p efficiently displaces Sub2p from Yra1p (Fig. 5c, lane 3). These competition data, together with the finding that Sub2p and Mex67p interact with the same regions of Yra1p, suggest that Sub2p and Mex67p/Mtr2p may bind to Yra1p sequentially. A possible implication of these data is that binding of the Mex67p/Mtr2p complex to Yra1p might release Sub2p from Yra1p before mRNA export.

To determine whether the interaction of Yra1p with Mex67p/

Mtr2p or with Sub2p is conserved, we used GST-tagged Aly, the metazoan homologue of Yra1p, for interaction assays (Fig. 5d). Significantly, GST–Aly binds to both Mex67p/Mtr2p and Sub2p (Fig. 5d, lanes 3 and 4). In contrast, GST–Npl3p, which was used as a negative control, does not bind to either protein (Fig. 5d, lanes 6 and 7). We conclude that the interaction between Yra1p and Sub2p, as well as the interaction between Yra1p and Mex67p/Mtr2p, is conserved.

We have shown that Yra1p, a conserved mRNA export factor, interacts with the conserved splicing factor Sub2p both genetically and physically. Thus, a splicing-coupled mechanism of mRNA export may also exist in *Saccharomyces cerevisiae*, in which about 40% of the total amount of transcripts derived from RNA polymerase II contain introns¹¹. In an accompanying paper¹⁸, evidence is provided that the mammalian homologue of Sub2p, UAP56, recruits the Yra1p counterpart Aly to the spliced mRNA, the complex that targets mRNA for export¹². Our findings show that Sub2p has a specific role in nuclear mRNA export. In addition, Sub2p and Mex67p compete for binding to Yra1p. Thus, Mex67p may replace Sub2p and then target the mRNP to the nuclear pores by direct interaction of the Mex67p/Mtr2p complex with nucleoporins. Together, our data suggest a model in which Sub2p is involved in recruiting Yra1p to mRNPs. Apparently, Sub2p functions in nuclear export of both intron-containing and intron-lacking mRNAs. There is a precedent for splicing factors being involved in nuclear export of mRNAs lacking introns in metazoans¹³. In this case, members of the SR protein family are recruited by a specific element in the exon. The mechanism by which Sub2p is recruited to mRNAs lacking introns remains to be determined. However, Aly was implicated in transcriptional co-activation¹⁴ and suggested to act as a nuclear protein chaperone that stimulates transcriptional activity of bZIP proteins¹⁵. Therefore, the interaction between Sub2p/UAP56 and Yra1p/Aly could likewise mediate the export of intron-lacking transcripts. □

Methods

Yeast strains

Yeast strains used in this study are provided in the Supplementary Information. Microbiological and yeast techniques were done essentially as described². Plasmids pUN100-*MEX67*, pUN100-*mex67-5*, pRS314-*YRA1* and L20GST-*YRA1* were reported previously¹². The *sub2Δ/yra1Δ* and *sub2Δ/mud2Δ* double-disruption strains were constructed by mating the *sub2* shuffle strain to the *yra1* shuffle or the *mud2* disruption strain, respectively, and dissecting the tetrads. The *sub2Δ/GAL::RRP41* and *mex67Δ/GAL::RRP41* double-mutant strains were constructed by mating the *sub2* or *mex67* shuffle strain to the *GAL::RRP41* strain⁹ and dissecting the tetrads.

Plasmid construction

Plasmids pRS314-*yra1-ΔN1* and pRS314-*yra1-ΔRRM* were constructed by mutagenic polymerase chain reaction (PCR). Plasmid pHT4467-*YRA1*, used for the synthetically lethal screen, was cloned by subcloning intron-lacking *YRA1* into plasmid pHT4467 (ref. 1). Plasmids L20GST-*YRA1-N*, -*M* and -*C* were cloned by amplification of the corresponding base pairs by PCR and insertion into vector L20GST. For recombinant, HIS-tagged Sub2p the open reading frame of *SUB2* was amplified by PCR and cloned into pET9d-HIS6-TB and expressed in BL21. The ProtA-TEV-tagged version of Sub2p was constructed by amplification of the *SUB2* open reading frame and cloning into pNOPATA1L (ref. 16) (*NOP1* promoter) or pGALPATG1L (*GAL1* promoter). The *sub2-85* mutant was obtained by mutagenic PCR as described².

Protein and mRNA analyses

For affinity purification of ProtA-TEV-Sub2p, cells were lysed with glass beads, proteins extracted with buffer (20 mM HEPES at pH 7.0, 100 mM potassium acetate, 2 mM magnesium acetate, 0.1% Tween 20, 5 mM β-mercaptoethanol, protease inhibitors), and the supernatant incubated with immunoglobulin-γ (IgG) Sepharose. The beads were washed with buffer, buffer containing 10 mM and 30 mM MgCl₂, and bound proteins eluted by cleavage with TEV protease in a final concentration of 100 mM MgCl₂. The synthetically lethal screen, the poly(A)⁺ *in situ* hybridization, protein expression in *S. pombe*, and the *in vitro* binding assay were performed as described². *In situ* hybridization using oligonucleotides specific for *SSA1* and *ACT1* mRNAs was performed according to ref. 7. Oligonucleotides specific for *SSA1* RNA were previously described⁷. *ACT1* mRNA was detected with a mixture of 5'-Cy3-labelled oligonucleotides as indicated in the Supplementary Information. SDS-PAGE and western blot analyses were performed as described¹⁷.

Received 11 June; accepted 21 August 2001.

1. Segref, A. *et al.* Mex67p which is an essential factor for nuclear mRNA export binds to both poly(A)⁺ RNA and nuclear pores. *EMBO J.* **16**, 3256–3271 (1997).
2. Sträßer, K. & Hurt, E. C. Yra1p, a conserved nuclear RNA binding protein, interacts directly with Mex67p and is required for mRNA export. *EMBO J.* **19**, 410–420 (2000).
3. Libri, D., Graziani, N., Saguez, C. & Boulay, J. Multiple roles for the yeast *SUB2/yUAP56* gene in splicing. *Genes Dev.* **15**, 36–41 (2001).
4. Zhang, M. & Green, M. R. Identification and characterization of yUAP/Sub2p, a yeast homolog of the essential human pre-mRNA splicing factor hUAP56. *Genes Dev.* **15**, 30–35 (2001).
5. Kistler, A. L. & Guthrie, C. Deletion of MUD2, the yeast homolog of U2AF65, can bypass the requirement for Sub2, an essential spliceosomal ATPase. *Genes Dev.* **15**, 42–49 (2001).
6. Portman, D. S., O'Connor, J. P. & Dreyfuss, G. Yra1, an essential *Saccharomyces cerevisiae* gene, encodes a novel nuclear protein with RNA annealing activity. *RNA* **3**, 527–537 (1997).
7. Hurt, E. *et al.* Mex67p mediates the nuclear export of a variety of pol II transcripts. *J. Biol. Chem.* **275**, 8361–8368 (2000).
8. Jensen, T. H., Patricio, K., McCarthy, T. & Rosbash, M. A block to mRNA nuclear export in *S. cerevisiae* leads to hyperadenylation of transcripts that accumulate at the site of transcription. *Mol. Cell* **7**, 887–898 (2001).
9. Bousquet-Antonelli, C., Presutti, C. & Tollervey, D. Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell* **102**, 765–775 (2000).
10. Stutz, F. *et al.* REF, an evolutionarily conserved family of hnRNP-like proteins, interacts with TAP/Mex67p and participates in mRNA nuclear export. *RNA* **6**, 638–650 (2000).
11. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24**, 437–440 (1999).
12. Zhou, Z. L. *et al.* The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature* **407**, 401–405 (2000).
13. Huang, Y. & Steitz, J. A. Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. *Mol. Cell* **7**, 899–905 (2001).
14. Bruhn, L., Munnerlyn, A. & Grosschedl, R. ALY, a context-dependent coactivator of LEF-1 and AML-1, is required for TCRα enhancer function. *Genes Dev.* **11**, 640–653 (1997).
15. Virbasius, C. M., Wagner, S. & Green, M. R. A human nuclear-localized chaperone that regulates dimerization, DNA binding, and transcriptional activity of bZIP proteins. *Mol. Cell* **4**, 219–228 (1999).
16. Hellmuth, K. *et al.* Yeast Los1p has properties of an exportin-like nucleocytoplasmic transport factor for tRNA. *Mol. Cell. Biol.* **18**, 6364–6386 (1998).
17. Siniosoglou, S. *et al.* A novel complex of nucleoporins, which includes Sec13p and a Sec13p homolog, is essential for normal nuclear pores. *Cell* **84**, 265–275 (1996).
18. Luo, M.-J. *et al.* Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly. *Nature* (this issue).

Supplementary information is available on *Nature's* World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We are grateful to R. Grosschedl for GST–Aly, J. Braspenning for L20GST–NPL3, J. Beggs for *prp4-1* and *prp16-2* mutants, and D. Tollervey for the strain *GAL10::ProtA-RRP41* (P118). We thank R. Reed and the members of the Reed lab for critical reading of the manuscript. E.H. was a recipient of grants from the Deutsche Forschungsgemeinschaft and Fonds der Chemischen Industrie.

Correspondence and requests for materials should be addressed to E.H. (e-mail: cg5@ix.urz.uni-heidelberg.de).

corrections

A mouse knock-in model exposes sequential proteolytic pathways that regulate p27^{Kip1} in G1 and S phase

Nisar P. Malek, Holly Sundberg, Seth McGrew, Kelko Nakayama, Themis R. Kyriakides & James M. Roberts

Nature **413**, 323–327 (2001).

In this Letter, the last name of Themis R. Kyriakides was misspelled as 'Kyriakidis'. □

CREB regulates hepatic gluconeogenesis through the coactivator PGC-1

Stephan Herzig, Fanxin Long, Ulupi S. Jhala, Susan Hedrick, Rebecca Quinn, Anton Bauer, Dorothea Rudolph, Gunther Schutz, Cliff Yoonll, Pere Puigserver, Bruce Spiegelman & Marc Montminy

Nature **413**, 179–183 (2001).

In the third sentence, “hyperglycaemia” should have read “hypoglycaemia”. The corrected sentence is: “Here we show that mice carrying a targeted disruption of the cyclic AMP (cAMP) response element binding (CREB) protein gene, or overexpressing a dominant-negative CREB inhibitor, exhibit fasting hypoglycaemia and reduced expression of gluconeogenic enzymes.” □

erratum

A titanasilicate molecular sieve with adjustable pores for size-selective adsorption of molecules

Steven M. Kuznicki, Valerie A. Bell, Sankar Nair, Hugh W. Hillhouse, Richard M. Jacobinas, Carola M. Braunbarth, Brian H. Toby & Michael Tsapatsis

Nature **412**, 720–724 (2001).

Correspondence and requests for materials relating to the structural analysis in this Letter should be addressed to Michael Tsapatsis (e-mail: tsapatsi@ecs.umass.edu). □

Shopping for gels

Electrophoresis equipment, software, buffers and the rest.

Gel Plate Rack

Bel-Art Products www.bel-art.com

Get a handle on this one

The Scienceware Electrophoresis Gel Plate Rack has 10 positions for holding glass plates, a weighted base for safety and wide carrying handles on both ends. It is made of smooth polypropylene material to cushion the glass edges, reducing chipping and cracking of plates. Labels can be affixed to a flat area on each end of the rack. The unit has a footprint of 8 × 15 inches and is 9 inches high.

Glass plate storage

C.B.S. Scientific www.cbsscientific.com

Hue and dry

C.B.S. have introduced 12-place storage/drying racks for handling electrophoresis gel plate. The racks are made of high-strength ABS and come in green, blue, gray and black. The colour-coding scheme is intended to make it easy for lab personnel to distinguish between clean and dirty, pre-treated and untreated, or front and back glass plates. There are hand supports to provide a secure grip when moving loaded racks around the lab.

CentriVap

Labconco www.labconco.com

Combining sample evaporation and gel drying

CentriVap gel dryer system enables users to concentrate samples and dry electrophoresis gels simultaneously using one cold trap and vacuum pump. Samples are rapidly evaporated using centrifugal force, heat and vacuum. Gels up to 33.6 × 42.3 cm in size can be dried. Separate vacuum and heat timers provide application flexibility. The concentrator and gel dryer are connected via a three-way valve so that both units may be used at the same time. The system includes the CentriVap Benchtop Concentrator, -55° cold trap, three-way valve and connecting tubing, with a diaphragm vacuum pump sold separately.

Slide-A-Lyzer

Pierce www.piercenet.com/salmini

Mini-dialysis units

Anyone struggling with homemade devices for small-volume dialysis may appreciate these units, supplied in packs of 50 or 250, or as a 10-unit trial pack that includes a newly designed multi-unit array float for dialysing 1–25 samples in organics, weak acids and



Two new offerings for plate storage: this 10-place one is from Scienceware, and C.B.S. do a '12' in colours.

weak bases. Slide-A-Lyzers are suitable for sample volumes as from small as 10 µl up to 100 µl. Sample recovery is 9–10 µl for a 10 µl sample, and a 100 µl, pH 2.8 buffer sample can be converted to pH 9.4 dialysing against a litre of bicarbonate buffer, pH 9.4 in less than 10 minutes. There is no need for a microcentrifuge or syringe adapters. Samples are pipetted into the unit, capped and dialysed. The units are made with low protein and nucleic acid-absorbing regenerated cellulose membranes in 3.5 K, 7 K or 10 K molecular weight cut-offs.

AlphaEaseFC 3.0

Alpha Innotech Corporation

www.alphainnotech.com

Easy-to-use software for gel imaging and analysis

Version 3.0 of AlphaEaseFC has been developed with a view to ease of use, and includes new features for automated image acquisition and analysis of fluorescence, chemiluminescence, and colorimetric gels, films, membranes, macroarrays, and microplates. New features include integrated 1D lane densitometry, molecular weight determination, Mass/QPCR standards, band scoring, and pattern matching. There is automatic exposure capability with user-defined settings for fluorescence, colorimetric, and

chemiluminescence samples and real-time display updates. AlphaEaseFC software is available as a stand-alone package that can be used with images captured from CCD, laser, or scanner based image acquisition systems and is also integrated into the FluorChem, ChemImager, and AlphaMager imaging systems manufactured by Alpha Innotech.

TTE sequencing buffer

AMRESCO Inc. www.amresco-inc.com

The long game

The latest addition to the SOLON-AMRESCO buffer range is Tris-TAPS-EDTA (TTE), a specially formulated buffer for use with the ABI 377 DNA Sequencer. TTE buffer is recommended for use in extended read slab-gel electrophoresis applications and is claimed to increase read lengths over standard TBE when used in automated sequencing. The Tris-TAPS-EDTA buffer is optimized to provide long read lengths and reproducibility, and is nuclease-free.

GeneDirectory

Syngene www.syngene.com

Software with a comparative advantage

GeneDirectory is a gel data storage and analysis software program that works in

conjunction with the company's GeneTools to compare thousands of bands on different gels. Data can be viewed on screen as familiar gel tracks. Up to five gel tracks containing DNA from known samples are shown in an upper window, while those bands on the gel that are to be compared are seen in a lower one, making the program particularly suitable for applications such as paternity testing. The software provides three ways in which to compare data, ranging from simple band comparison to more complex co-dominance and dominance analysis suitable for RFLP and AFLP studies, forensic work or gene mapping, the results can be exported to Excel spreadsheets or text files.

Slimline TE-312S

Spectronics Corporation

www.spectroline.com

Spotlight on mini-gels

Although smaller than most laptops, this compact UV transilluminator for visualization and photo-documentation of mini-gels is claimed to offer most of the features of larger units. It measures $29.2 \times 35.6 \times 5.7$ cm and weighs just 5.4 kg. Models with 120V/60Hz, 230V/50Hz and 100V/50-60Hz are available. The unit produces a typical peak 312 nm intensity of 9,000 mW per cm^2 at the filter surface. A diffusing screen ensures superior light distribution and eliminates confusing light striations, which are caused by the contours of the tubes. The Longlife filter glass on the transilluminator inhibits solarization for up to 50 times longer than ordinary filters, according to the manufacturer, resulting in longer filter life. The unit features a hinged UV-blocking cover to shield the user from UV radiation and protect the filter from damage caused by dropping objects onto the glass. Transmission of 254 nm radiation is eliminated, minimising photobleaching, photonicing and photodimerization.

InstaSTAIN EtBr

Trevigen

www.trevigen.com

More haste, less waste

The InstaSTAIN Et Br DNA gel staining agent comes in sheet form, so that a precise amount of dye can be delivered direct to DNA gels, eliminating the need to dissolve ethidium bromide in large volumes of gel and/or running buffer. Using a microwave protocol, staining is performed in five seconds. The product is designed to reduce ethidium bromide biohazardous waste.

AccuGENE

BioWhittaker Molecular

www.bmaproducts.com

Ready-to-use electrophoresis buffers

BioWhittaker offers TBE, TAE, CE, Tris-glycine and MOPS buffers formulated to

match its pre-cast Reliant, Latitude and PAGEr Gold gels and Long Ranger sequencing products. The buffers use the same reagents as those used in the pre-cast gels and sequencing solutions, resulting in a closed system designed to yield reproducible results.

Hoefer SE 600 Ruby

Amersham Pharmacia Biotech

www.apbiotech.com

New-look electrophoresis unit

An ergonomically designed version of the Hoefer standard vertical slab electrophoresis unit, the SE 600 Ruby can be used to run single or multiple gels simultaneously by using two gels in the basic unit or four gels with accessory divider plates. A total of 112 samples can be separated in a single run using four gels. The unit maintains temperatures of between 1°C and 48°C and a built-in heat exchanger allows fast runs with no loss of band separation. The unit is fully compatible with the existing Hoefer SE 600 model accessories and components. Applications include SDS gel and second-dimension separation of two-dimensional electrophoresis.

SYPRO

Cambridge Bioscience www.bioscience.co.uk

There's no stopping this stain

SYPRO ruby is a ready-to-use, luminescent stain for the detection of proteins separated using one- or two-dimensional polyacrylamide gel electrophoresis. It offers a sensitivity that is comparable to that of silver stains, but unlike them uses a rapid, one-step protocol that can be completed in 90 minutes. No stop solutions or de-staining procedures are required. The stain produces linear quantification over three orders of magnitude with less protein-to-protein variability than silver stains. It also works with proteins that can be difficult to stain, including glycoproteins and lipoproteins, and does not stain extraneous nucleic acids or interfere with

subsequent analysis of proteins by Edman-based sequencing or mass spectrometry. SYPRO Ruby is supplied in 200 ml, 500 ml and 1 litre sizes.

Phoretix 1D v 5.0

Nonlinear dynamics

www.nonlinear.com

Software for high-throughput one-dimensional electrophoresis

Version 5 of Phoretix 1D contains a lane and band detection algorithm capable of dealing with a broad range of 1D electrophoresis applications. The entire gel analysis process can be automated through a customised list of operations, running in a set order to produce a complete analysis — from a 1D gel image to a comprehensive set of experimental data. Although highly automated, the program offers the user analytical flexibility and control for checking results step by step and an intuitive, easy to use interface. Applications include analysis of multi-tier gels, MADGE gels and large, multi-well format, sequencing style gels.

LOAD'n GLO

PanVera Corporation

www.panvera.com

Labelled DNA markers

The LOAD'n GLO set of labelled DNA size markers has been developed primarily for use in non-radioactive Southern blot applications. Markers are available labelled with biotin, digoxin, dinitrophenyl or fluorescein. They are formulated to allow visualization through ethidium bromide staining of the gel or enzymatic detection of the specific epitope after blotting to a membrane. Each marker is supplied at a concentration of 0.25 μg per μl with sufficient product for 50 lanes. A loading volume of 2 μl per lane is recommended. All markers are supplied in the following storage buffer: 10 mM Tris (pH 7.5), 1 mM EDTA, 10 mM NaCl.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

**DIVERSE
SMALL
MOLECULES
READY FOR
SCREENING**

**10K-50K
Compounds
in Plates**

• Next Day Delivery*
• Competitively Priced

*From New York to Europe

Cambridge Corporation • 800.964.6100
552.42.3400 • 552.42.3401 • E-mail: chem@cambridge.com

www.cambridge.com

Look what's new...

EDG (GPCR)

- Human Recombinant GPCRs
- Membrane Preps
- Lysophospholipid growth factor receptors

...From Exalpha Biologicals, Inc.

Find out more
Phone: 617-443-4545 x 304
Website: www.exalpha.com

Exalpha

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Group European Manager:

Leonie Wells (4954)

European Manager:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Silje Opstrup (4994)

Spain/ Portugal: Leonie Wells (4954)

Production Manager:

Billie Franklin

To send films and materials use

London address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau

Tel +33 (0) 1 43 20 16 51

Fax +33 (0) 1 43 20 51 52

e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner

Tel +49 89 54 90 57 11/-2

Fax +49 89 54 90 57 20

e-mails: p.phelan@nature.com

k.turner@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

MG Ichigaya Building (6F),

19-1 Haraikatanmachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: k.cowan@naturejp.com

Japan Manager:

Kate Cowan

naturejobs

nature.com

California, home of laureates

How do you recruit Nobel laureates to your region? It helps to have a concentration of them already in place. But other factors, such as climate, culture and commitment to excellence play a role. California, which excels at both producing laureates and attracting them from other parts of the world, will this month join in the celebrations of the award's centenary.

The state has an enviable association with the Nobel prize. It has the highest concentration of laureates in the world — nearly 25% of laureates in the past 10 years claim a California connection.

One of them, David Baltimore, president of the California Institute of Technology, has some theories on why the state has been both a magnet and an incubator for laureates. Baltimore, who shared the 1975 prize in physiology or medicine for his contributions to showing how viral reverse transcriptase can lead to cancer, says that a concentration of excellent public and private research universities has helped.

But attitude — more than location and climate — has made those universities conducive to the kind of science that eventually gets rewarded. California's universities, born early in the last century, are "unfettered by historical commitments and hierarchical formality", Baltimore says.

Other regions throughout the world are working on the easy part: assembling a critical mass of universities, government labs and research-based businesses within close proximity. The less tangible necessity — not only creating an environment that fosters an attitude of innovation, but keeping that spirit as the place becomes established — is much more difficult. Having laureates on staff does help to attract money and talent, Baltimore says. But, he warns: "Resting on your laurels is a poor idea."

Paul Smaglik
Naturejobs editor



Contents

SPECIAL REPORT

US physics graduates prepare for an industrial future **p4**

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific
job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



Across America, universities are now expanding their courses to prepare physics graduates for careers in industry, says Steve Bunk.

Forward thinking: Jim Freericks puts his students through a range of unusual exercises in the physics for industry course.



In a classroom at Georgetown University in Washington DC, four PhD candidates in their early twenties uncertainly face associate physics professor Jim Freericks. The students are nervous because they suspect that they'll be asked to do strange and awkward things.

Freericks asks the three women and a man to stand back-to-back in pairs, each person holding a sheet of paper. Two postdocs and two professors join the exercise, for a total of four pairs. One person in each pair must fold the paper four times, at least once diagonally, explaining every fold in detail to the other team member, who must follow directions without speaking. When the pairs turn around, all four results are utterly mismatched. Laughing, they try again and do slightly better, although one postdoc is finally moved to tell his partner: "Just make a paper airplane and hope it flies."

Although this exercise in blind communication has nothing to do with physics, it and other similar exercises (see 'Lateral thinking required', opposite) serve nicely as an introduction to Georgetown's programme on industrial leadership in physics. Launched this autumn after five years of planning, this programme aims to prepare students for careers in the private sector.

A final test on this orientation day does involve physics — calculating the trajectory to launch a spring accurately into a box — but even this is couched in the language of team solutions, by assignment of a specific role (manager, energizer, summarizer, sceptic) to each participant. All of these exercises are primers for the new curriculum's 'integrative experiences', in which student teams work on well-defined projects with tight deadlines. The coursework is largely condensed-matter physics, which has a broader range of business applications than, say, high-energy physics. There are also internships with companies.

INDUSTRY TRAINING SOARS

Over the past few years, such business-related graduate physics programmes have become increasingly popular. More than 60 universities across the United States now offer courses that prepare physics graduates

for careers in industry rather than academia. Many of these programmes include internships with companies, most have only a handful of students per year, and if their widely differing curricula have anything in common, it is probably that they explore the dynamics of solving problems in groups. That, and one other thing: all but the Georgetown programme terminate in a master's degree.

Freericks says that the Georgetown programme was inspired in part by a "frightening reduction" in physics undergraduate and graduate students since the end of the cold war a decade ago, which brought an influx of talented physicists to America. Although that influx has since been absorbed, educators and other specialists argue that the perception of a tight physics job market lingers among students. The reality is a bullish job market, with much of the growth in industry, according to statistics from the American Institute of Physics. But the ascendancy of biology and its related computational sciences, such as bioinformatics, has drawn many of the best and brightest American students away from physics. Data from the National Science Foundation show that in 1999–2000, foreign first-year physics graduate students were in the majority for the first time in history.

Author and science-education consultant Sheila Tobias notes that about 3,800 students gain a BS in physics from US universities annually, compared with about 64,000 in biology. "The overarching goal is to make physics a mainstream subject," she says. In 1995, Tobias and two colleagues called for professional master's degrees (PMDs) in science. "We were three lone voices," she says, although in the same year a National Academy of Sciences study broached the idea that science graduate programmes should consider creating degrees orientated less towards research and more towards industry.

In 1997, two foundations mounted separate initiatives. The W. M. Keck Foundation of Los Angeles began to set up the Keck Graduate Institute in Claremont, California, which now recruits undergraduates from physics and other science programmes for master's training in both the business and science aspects of employment in industry. Meanwhile, the Alfred P. Sloan Foundation of New York started arranging its current funding of 46 PMD programmes at 27 universities, of which four are in physics. Another 57 physics PMDs are available, a

Lateral thinking required

Jim Freericks at Georgetown University in Washington asks his students to imagine having crash-landed in a subarctic country in autumn, 22 air miles from the nearest town. Fifteen things have been salvaged from the plane, and each person must rank the items in order of their importance to survival.

After the students finish their lists, they are asked to discuss and decide by consensus on a joint ranking for the items. Finally, Freericks compares the rankings to those of experts. Although the students tend to perform better

as a group, they can make what the expert survivalists describe as a fatal decision to go towards town rather than wait for help. In fact, the students don't carefully consider what to do, only deciding indirectly upon the trek as they discuss the potential use of each item.

Freericks then tells them about the Abilene paradox, in which a family drives a long way in bad weather for a dinner nobody wants, because no one can say "Let's stay home". He asks if the students have a perception of who was their leader — but they say it was teamwork. S.B.

recent American Physics Institute study reveals.

"The word 'professional' is very intentional," says Tobias, who is coordinator of the Sloan initiative. "We're trying to broaden the training, tasks and responsibilities of people trained in science." Success in the market-place requires, in part, creating cachet for master's programmes, which are often considered consolation prizes for those who do not achieve doctorates.

"There's a lot of confusion about what the correct level is for a master's degree," says physics professor Hans Bozler, who coordinates the Sloan-funded physics for business applications PMD programme at the University of Southern California. Some PMD programmes last only nine months and contain one or two specialized physics courses that are not "mainstreamed" into the normal graduate-school curriculum, he explains. Others extend to the full two years and have stringent graduation requirements. "They both carry the label of master's, and that's unfortunate. We see our competition as business schools and law schools," he says.

COURTING COMPANIES

The ease of securing participation by the private sector seems to vary according to factors such as personal contacts and regional needs. "I've got more companies at this point than I have students, but not every student is right for every company," says physics professor Cyrus Taylor, who directs the physics entrepreneurship programme at Case Western Reserve University (CWRU). This "lean and flexible" agenda focuses on innovation, he says. Ten students admitted in the programme's first two years are either perfecting physics and business models with the aim of starting up a company, or are solving problems of innovation for existing companies where they are interns.

The CWRU programme arose from entrepreneurial seminars funded by an alumnus grant intended to help science students learn about business without total reliance on the 'school of hard knocks'. Taylor thinks that his curriculum's emphasis on real-world involvement is less psychologically conservative than that of a traditional PhD programme, because it doesn't fear failure. "In a start-up, failure and adversity and risk are part of it. Not everything is going to succeed, and that is one of the important lessons to

learn," he says. "Also, the notion of what constitutes acceptable research, and how it should be funded, and where interesting problems come from is one that can be revisited, I think, from this kind of contact."

Proponents of PMDs say that they want to supplement rather than replace PhDs in the private sector. Freericks staunchly defends the duration and breadth of Georgetown's five-year doctoral programme, which includes 16 months of coursework, a year-long industrial apprenticeship and 30 months of dissertation research. "PhDs will always be employed by industry," he maintains. "I think you'll find more innovations come from people who have PhDs than people who don't, and it really comes from doing research."

Yet, for masters of physics who do not do research, industry can offer much more than dead-end technician jobs, says Tobias. Challenging careers in such fields as electronics, optics and computational modelling are open to them, along with opportunities in intellectual property law, continuing education and entrepreneurship.

As Bozler puts it: "What we're really interested in is finding students who eventually will take leadership roles in business."

Steve Bunk is a science writer based in Alexandria, Virginia.

Web links

Professional science masters

♦ www.sciencemasters.com

AIP physics masters study

♦ www.aip.org/professionalmasters/profmshigh.htm

Georgetown physics programme

♦ magus.physics.georgetown.edu/graduate.htm

Keck Graduate Institute

♦ www.kgi.edu